Jussi Ekström

Statistical Analysis of Large Scale Wind Power Generation

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 28.1.2014

Thesis supervisor:

Prof. Matti Lehtonen

Thesis advisors:

M.Sc (Tech.) Jussi Matilainen

M.Sc. (Tech.) Matti Koivisto

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Jussi Ekström

Title: Statistical Analysis of Large Scale Wind Power Generation

Date: 28.1.2014 Language: English Number of pages: 8+72

Department of Electrical Engineering

Professorship: Power systems Code: S-18

Supervisor: Prof. Matti Lehtonen

Advisors: M.Sc (Tech.) Jussi Matilainen, M.Sc. (Tech.) Matti Koivisto

The amount of wind power generation is increasing in many countries and therefore, the effects of wind power on the electric power system are becoming more and more important. Two time series models are developed in this thesis, the transformed VAR model with time-dependent intercept term and the transformed ARC model. The models can be used in Monte Carlo simulations to determine the risk of very high or low wind speeds occurring contemporaneously in several locations. The feasibility of the models is evaluated for existing measured locations and new non-measured locations. The models are verified against 21 measurement locations from Finland. In addition, example scenarios of the applications of the models are presented.

Keywords: Copula, Monte Carlo simulation, Vector autoregressive model, Weibull distribution, Wind speed, Wind power

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Jussi Ekström

Työn nimi: Laajamittaisen tuulivoimatuotannon tilastollinen analyysi

Päivämäärä: 28.1.2014 Kieli: Englanti Sivumäärä: 8+72

Sähkötekniikan laitos

Professuuri: Sähköjärjestelmät Koodi: S-18

Valvoja: Prof. Matti Lehtonen

Ohjaajat: DI Jussi Matilainen, DI Matti Koivisto

Tuulituotannon määrä kasvaa jatkuvasti monissa maissa, ja täten lisääntyneen tuulituotannon vaikutukset sähköjärjestelmään tulevat yhä merkittävämmiksi. Tässä diplomityössä kehitettiin kaksi aikasarjamallia, VAR-malli aikariippuvalla leikkaustermillä muunnetulle datalle ja ARC-malli muunnetulle datalle. Malleja voidaan käyttää Monte Carlo -simulaatioissa sellaisten tilanteiden todennäköisyyksien määrittämiseen, missä esiintyy erittäin suuria tai matalia tuulennopeuksia samanaikaisesti monissa eri kohteissa. Mallien käyttökelpoisuutta arvioidaan kahdenlaisissa tilanteissa, sellaisissa joissa mallinnetaan olemassa olevia kohteita joista on mittausdataa, ja sellaisissa, joissa mallinnetaan uusia kohteita joista ei ole lainkaan mittausdataa. Esitetyt mallit todennetaan vertaamalla niiden antamia simulaatiotuloksia 21:een mittauskohteeseen Suomesta. Lisäksi esitellään esimerkkitilanteita mallien eri sovellusmahdollisuuksista.

Avainsanat: Autoregressiivinen malli, Kopula, Monte Carlo -simulaatio, Tuulen nopeus, Tuulivoima, Weibull-jakauma

Preface

I want to express my gratitude to my supervisor, Professor Matti Lehtonen, for his help, comments and him taking the position of the thesis supervisor on such a short notice.

I express my gratitude also to my thesis instructor Matti Koivisto for his support and invaluable ideas and comments.

I want to thank my instructor Jussi Matilainen from Fingrid Oyj for his comments and ideas and Fingrid Oyj for the funding of my thesis.

I would also like to thank everyone in the steering group of my thesis and the research involved. Thank you Professor Liisa Haarla, Ilkka Mellin and Janne Seppänen for ideas and comments.

My parents also deserve thanks for their continuous help and support not to forget the tremendous encouragement to academic pursuits.

To Sanni, for her love and support.

Otaniemi, 28.1.2014

Jussi Ekström

Contents

Nomenclature

Symbols

$\boldsymbol{A}$	coefficient matrix of vector autoregressive models
A	Weibull scale parameter
a	autoregressive model parameter
B	Weibull shape parameter
$\boldsymbol{C}$	correlation matrix
C	copula
c	intercept term
D	dummy variable
F	cumulative distribution function
F^{-1}	inverse cumulative distribution function
h	lag
$\boldsymbol{I}$	identity matrix
p	order of autoregressive model
$\boldsymbol{R}$	autocorrelation matrix
r	Spearman's rank-order correlation coefficient
t	time
u	error term
μ	mean
ρ	Pearson's correlation coefficient
$\boldsymbol{\Sigma}$	covariance matrix
σ	standard deviation
σ^2	variance

Abbreviations

ACF	autocorrelation function
ARC model	univariate autoregressive model with Cholesky decomposition
ARMA model	autoregressive-moving-average model
AR model	autoregressive model
CCOV	cross-covariance function
CDF	cumulative distribution function
E	expected value
ECDF	empirical cumulative distribution function
ET	Energiateollisuus
FMI	The Finnish Meteorological Institute
GARCH model	Generalized autoregressive conditional heteroskedasticity model
ML	maximum likelihood
OLS	ordinary least squares
PDF	probability density function
RACF	ranked autocorrelation function
RXCF	ranked cross-correlation function
Var	variance
VAR model	vector autoregressive model
VARX model	vector autoregressive model with exogenous variables
XCF	cross-correlation function

1 Introduction

The amount of wind power generation in the total power generation structure is growing rapidly in many countries, including Finland. The European Union climate and energy package published three 20 targets in 2007. The targets include both to reduce the emissions of greenhouse gases by 20 percent to 2020 and to reach 20 percent of renewable energy in total energy consumption in the EU by 2020. The recently published targets for 2030 also encourage and direct more to the increase in the renewable energy generation. These targets indicate that the amount of wind power generation will increase also in the future, if these goals are to be achieved.

This rises new challenges because of the variable nature of the wind speeds and thus wind power respectively. The wind generation is varying by nature, which creates uncertainty also in the power generation in general. The uncertainty in generation, caused by the intermittent nature of wind power, has to be taken into account in the planning and operation of power systems with a growing number of wind turbines. Also, the effect of the increasing amount of wind power generation has to be considered in the long-term grid planning. Therefore, it is necessary to understand and to be able to assess the behaviour of wind power generation in several locations contemporaneously.

The wind power generation uncertainties can be analyzed over both short and long time scales. Short-term analysis focuses on the operation of the power system and long-term analysis to the planning of the power system. The short-term analysis can be beneficial to, for example, the power system operator and the long-term analysis for the use of long-term power system planning. This thesis focuses on the long-term analyses of wind power generation and presents two Monte Carlo simulation models, which are able to assess different wind power generation scenarios.

The main objective of this thesis is to develop two statistical models for the assessment of large scale wind power generation and evaluate the feasibility of the models. The models should be able to produce as accurate simulation as possible with the existing data and existing wind generation locations, and simulations of scenarios with new non-measured wind generation locations. The developed models are able to assess the probabilities for scenarios where the wind power generation is very high or very low in the system. The models can be useful tools, for example, for the assessment of the required power generation reserves and long-term network planning studies in power systems with a large penetration of wind power generation.

A transformed vector autoregressive (VAR) model with a time-dependent intercept term and a univariate autoregressive model with correlations obtained with Cholesky decomposition (transformed ARC model) are developed and presented. The transformed VAR model with time-dependent intercept term is able to capture the structure of the wind speed data more precisely than the transformed ARC model, but it can be used only when simulating scenarios with existing measured locations. The transformed ARC model is also able to produce good results compared

with the transformed VAR model with time-dependent intercept term and the main advantage of this model is that it can be used to simulate scenarios with arbitrary non-measured locations implemented.

The structure of the thesis is following. Chapter 2 presents a short introduction about the subject. It considers relevant references for this thesis and presents the current wind power generation structure of Finland. Chapter 3 introduces the relevant theoretical background required including probability distributions, copula theory and autoregressive models used in this thesis. Chapter 4 presents the different measurement data used with the models. Chapter 5 introduces the full simulation models applied to wind speed simulations. The construction of the transformed VAR model with time-dependent intercept term and the transformed ARC model is presented. Chapter 6 presents the verification of the transformed VAR model with time-dependent intercept term and the transformed ARC model against the measurement data. Both models are verified against existing locations and transformed ARC model also against the addition of new non-measured locations. Chapter 7 presents simulations of different example scenarios with new locations and analyses of the results of these simulations. Chapter 8 discusses the results obtained in this thesis and Chapter 9 concludes the thesis.

2 Background of the Thesis

This chapter presents the current situation in the field of statistical analysis of large scale wind power generation, a closer look at the contents of this thesis and the current status of wind power generation in Finland. The relevant publications, which form the basis for this thesis, are presented and linked to the contents of the thesis and the wind power generation structure of Finland is considered.

2.1 Literature Review

Wind speeds are commonly considered as Weibull distributed. This is also the case in this thesis as the marginal distributions of the wind speeds are considered as Weibull distributions. Weibull distribution is beneficial to use because the Weibull parameters for different altitudes for every location in Finland are available in the Wind Atlas Database, which is a web portal hosted by The Finnish Meteorological Institute [1]. As wind speed conditions are known in multiple altitudes in every location in Finland, wind turbines of a different kind can be placed in arbitrary locations in the simulated scenarios.

The empirical cumulative distribution functions (ECDFs) for wind speeds have been considered by Klöckl [2] and Xie et al. [3], but they are not implemented in the analyses in this thesis since they are unable to estimate any events that are not present in the data. In case of ECDFs if the measurement data has the biggest measured wind speed at, for example 20 m/s, ECDFs will give zero probability for any event where the wind speed would be larger than 20 m/s. Therefore, the analysis of the probabilities of events not present in the data is not possible with ECDFs as stated by Nyström et al. [4] These events are crucial in the planning and operation of power systems and therefore, the ECDFs are not considered as a feasible option compared with fitted probability distributions like Weibull distribution.

In this thesis, copulas are used for transforming the measurement data from wind speeds to normal distribution and back in the models presented. The copula method has been used in the analysis of wind power generation in several locations by Xie et al. [3], Coić et al. [5], Louie [6] and Papaefthymiou et al. [7]. The most relevant advantage of the copula modeling is that the dependence structure analysis of several locations can be done separately from the analysis of the marginal distributions of the individual locations. Without copula modeling, the analysis of the marginal distributions and the dependency structure would together be very difficult.

There is a large number of different copulas, which can be used in the modeling. Xie et al. [3], Coić et al. [5], Louie [6], Stephen et al. [8] and He et al. [9] introduce the most common copulas, which are Gaussian, Archimedean and t-copulas. In this thesis, mainly the Gaussian copula is considered as it is the only copula, which can be used with the transformed ARC model and the transformed VAR model with time-dependent intercept term. Also, Archimedean copulas and t-copula are shortly

discussed in Section 8.1 as it is theoretically possible that the dependency structures specified by the presented models could be defined by other copula than Gaussian copula.

Both, the spatial and temporal dependency structures have to be analysed to obtain a simulation model that can capture all of the important characteristics of the uncertainties in the wind power generation. Autoregressive models have been used in wind speed simulations by Brown et al. [10]. The simulated AR(1) time series have been used with the Cholesky decomposition by Villanueva et al. [11]. The time series are multiplied by the Cholesky decomposition of the correlation matrix. With this procedure, cross-correlated time series can be obtained. This method is not able to capture the full spatial and temporal dependency structure, but it allows a straightforward implementation of new non-measured locations. According to Bechrakis et al. [12], to capture the full dependency structure, a more thorough consideration of the cross-correlation functions (XCFs) between locations is required. This can be done with a vector autoregressive model (VAR model), which has been applied by Klöckl [2], Hill et al. [13] and Klöckl et al. [14].

A VAR model has been introduced by Klöckl [2] and Klöckl et al. [14] and a more complex VARTA model by Deler et al. [15]. The transformed VAR model presented in this paper is based on a combination of these two models. In [15], Deler et al. had difficulties in the determination of the Pearson's correlations after the transformations of the data. This thesis shows that these difficulties can be avoided by defining the correlations as Spearman's rank-order correlations instead of Pearson's correlations as done by Xie et al. [3], Papaefthymiou et al. [7], [16] and Klöckl [14].

The dependency structure of several different locations is often measured with the autocorrelation function (ACF) and cross-correlation function (XCF). As mentioned, this paper applies the copula modeling and therefore the rank-preserving Spearman's rank-order correlation is used instead of Pearson's linear correlation. Therefore, in this thesis, a Spearman's rank-order autocorrelation function (RACF) and cross-correlation function (RXCF) are used to measure the dependence structures.

The analysis of the changing day structures i.e. the monthly diurnal variations is a vital part of wind speed modeling as shown by Klöckl [2], Brown et al. [10], Hill et al. [13] and Klöckl et al. [14]. There are numerous different options to analyze the diurnal structures. Next, few relevant options for the analysis are presented.

The diurnal structure has been analyzed by fitting multiple distribution to each location by Klöckl [2] and Klöckl et al. [14]. This procedure enables that each hour of the day in each different month has a different marginal distribution of wind speeds, which causes the required variation in the averages. This approach was not considered feasible with the data used in this thesis. Instead, for the transformed VAR model with time-dependent intercept term the diurnal structures are considered by extending a transformed VAR model with a time-dependent intercept term. This time-dependent intercept term captures the monthly diurnal variation in the

averages with a dummy variable system [17], [18]. This extended model allows the analysis of the full dependence structure of the data on a single vector autoregressive model with exogenous variables (VARX) estimation process. The full dependence structure includes both spatial and temporal dependencies and the monthly changing diurnal structures. The more detailed explanation of the analysis of the diurnal structures in case of the transformed VAR model with time-dependent intercept term can be found in Section 5.2.1.

Another approach to the analysis of the day structures is to remove the day structure from the data before the estimation of the model parameters and adding it back in the simulation phase as done by Hill et al. [13]. This method has been used with the transformed ARC model presented in this thesis and a more detailed explanation of the analysis can be found in Section 5.3.1.

2.2 Wind Power Generation and Capacity in Finland

The wind power capacity in Finland has increased rapidly in last few years. The first wind turbine in Finland started operation in 1991. Figure 1 shows the increase in the capacity from 1991 to the end of 2012. As seen in Figure 1, the capacity at the end of 2012 was approximately 288 MW. The capacity data has been acquired from Energiateollisuus (ET). The data is available for everyone free of charge but it has to requested from ET. Since 2013 generation data is not available without notable costs, we consider the situation at the end of 2012.

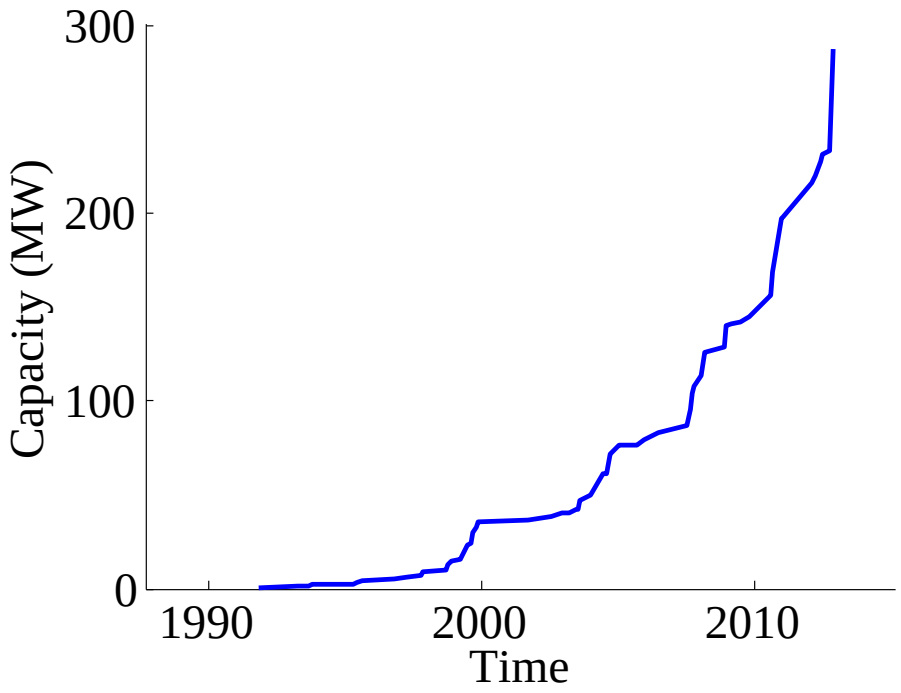

Figure 1: The wind power capacity in Finland from 1991 to the end of 2012.

The aggregated wind power generation i.e. the hourly generated energy can be found in the Figure 2. It is visible that the realized generation is significantly lower than the installed capacity at all times. It can be observed that the actual generation is never equal or even close to the installed capacity, which can be a result from many different factors. Some of those factors could be overall low wind speed conditions in the generation sites and the fact that a high geographical spread reduces the amount of extreme cases where there is no generation at all or the generation equals the installed capacity. Also, problems with the use of the turbines can reduce the hourly generated energy.

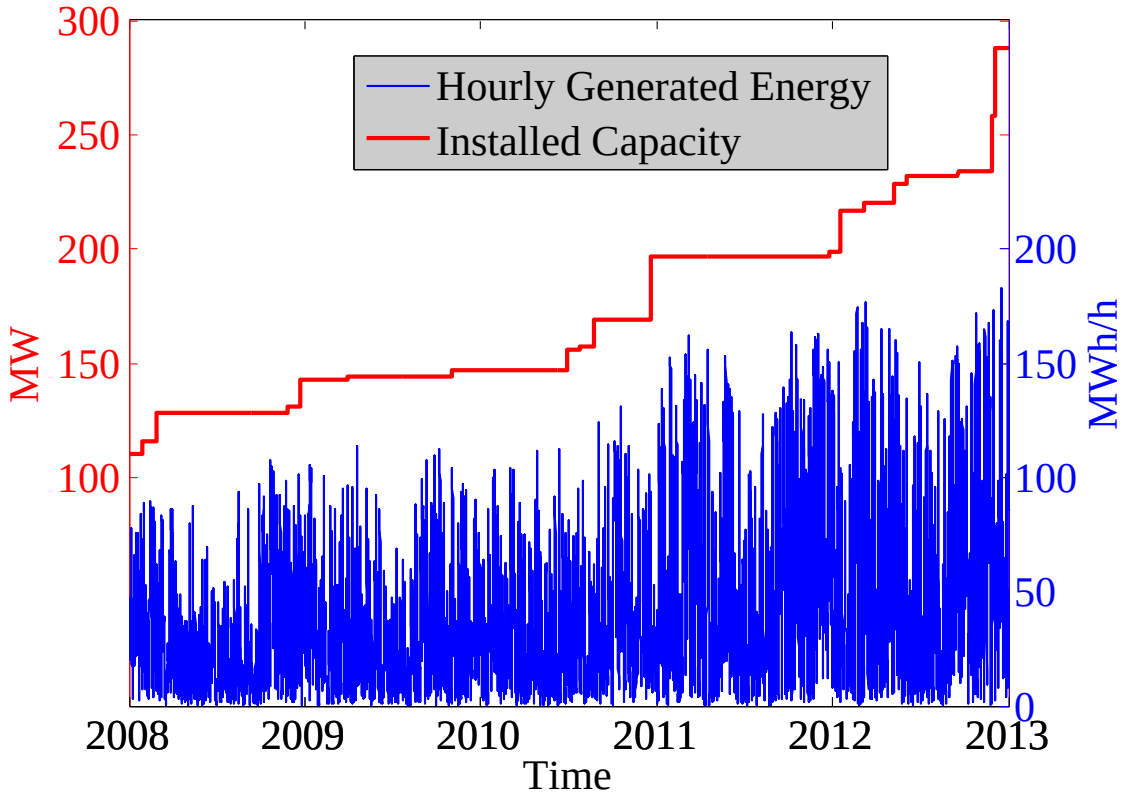

Figure 2: The hourly generated energy and total installed capacity in Finland from 2008 to the end of 2012. The data can be acquired from Energiateollisuus (ET) through request.

The locations of the individual turbines can be found in the Figure 3. The location data is freely available and it is maintained by VTT [19]. As visible in the Figure 3, the turbines are spread notably in the North-South axis as they can be found in the southwestern archipelago of Finland to the most northern parts of the country. On the West-East direction, the geographical spread is not as significant as the most of the turbines are located in the western coast of Finland. There are multiple wind farm projects currently going on in the eastern parts of Finland, which will increase

the geographical spread of the generation to West-East-direction.

The aggregate power generation data presented in Figure 2 is used in Chapter 7 of this thesis. The example cases presented in Chapter 7 are compared against the generation data from Finland between 2008 and 2012 when the feasibility of the cases is assessed.

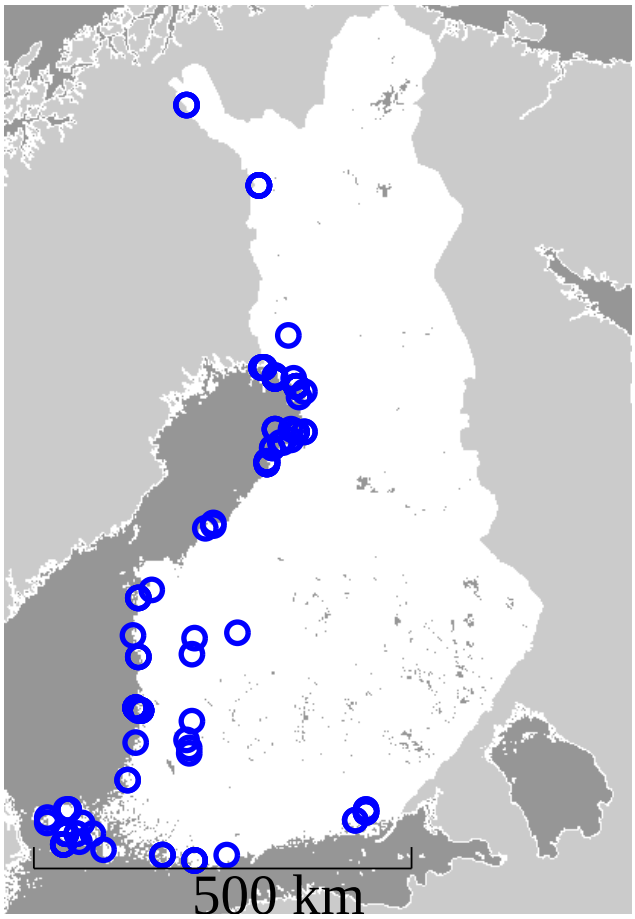

Figure 3: The locations of the installed wind turbines in Finland at the end of 2012. The location data is freely available and it is maintained by VTT. [19].

3 Theory of the Time Series Models

This chapter considers the required theoretical background to the time series models presented in this thesis. First, the relevant univariate and multivariate probability distributions are introduced. After that, the copula theory follows as the copula transformation is a crucial part of both of the models introduced in this thesis. The fourth part of this chapter focuses on the different autoregressive models and the most relevant models, in the viewpoint of this thesis, are presented. These models are an univariate autoregressive model (AR-model), a vector autoregressive model (VAR-model) and a vector autoregressive model with exogenous variables (VARX-model). Also, the estimation of the model parameters is included in this chapter. At the end of this chapter, the transformed VAR model and the transformed ARC model are introduced. These models are used as a basis for the construction of the wind speed simulation models in Chapter 5.

3.1 Univariate Probability Distributions

Probability distribution is a statistical concept that describes the probability that a random variable, drawn from a certain probability distribution, can fall into a certain interval or get a specific value. It is a function that tells a probability for a number (a random variable) to have a value between any two real numbers. Every sample is drawn from a probability distribution that contains the information concerning that specific random variable.

Probability distributions can be divided into continuous and discrete distributions. If a random variable is discrete, for example a value of a throw of a dice, the value is drawn from a discrete distribution. In case of continuous variables, they are drawn from a continuous probability distribution. This thesis considers only continuous distributions, so discrete distributions are not discussed further.

Probability distributions can be depicted with cumulative distribution functions (CDFs) and probability density functions (PDFs). CDF describes the probability that a random variable with a certain probability distribution will have a value that is less than or equal to an arbitrary value x. PDF is a derivative of the CDF, it is also nonnegative and its integral over the domain of the distribution is equal to one.

A probability distribution can be univariate or multivariate. Univariate distribution gives the probabilities of a single random variable and a multivariate distribution gives analogously the probabilities of a random vector, which contains two or more random variables. Multivariate distributions are considered in Section 3.2.

Next, the relevant univariate probability distributions for the topic of this thesis are shortly introduced.

3.1.1 Normal Distribution

Normal distribution is a common continuous probability distribution, also known as a Gaussian distribution and it is commonly denoted by $N(\mu, \sigma^2)$, where μ is mean and σ^2 is variance of the distribution.

The probability density function of a normal distribution is defined as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{1}$$

where μ is mean of the distribution and σ is the standard deviation. When $\mu = 0$ and $\sigma = 1$, the distribution is called a standard normal distribution. The probability density function of a standard normal distribution can be seen in Figure 4.

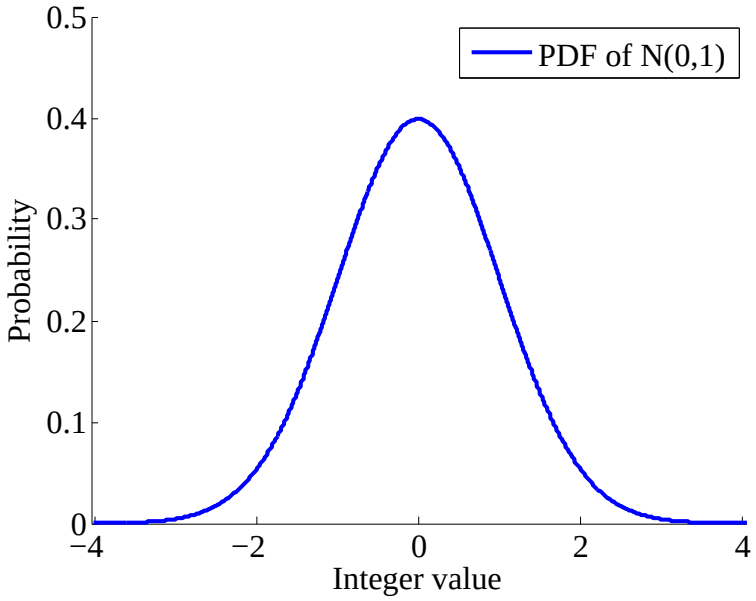

Figure 4: The probability density function of a standard normal distribution.

3.1.2 Weibull Distribution

Weibull distribution is a continuous and nonzero probability distribution and its probability density function for a random variable x can be written as

$$F(x; A, B) = \begin{cases} \frac{B}{A}\left(\frac{x}{A}\right)^{B-1} e^{-(x/A)^B} & x \geq 0, \\ 0 & x < 0, \end{cases} \tag{2}$$

where $A > 0$ is the scale parameter and $B > 0$ is the shape parameter of the Weibull distribution. The shape of the Weibull distribution depends on parameters A and B and Figure 5 illustrates the shape of the distribution with different scale parameters

when the shape parameter is fixed to value of 2. Weibull distribution has many uses and it has been introduced because it can be used to describe wind speed distributions. In this thesis, the wind speed is considered to be Weibull distributed.

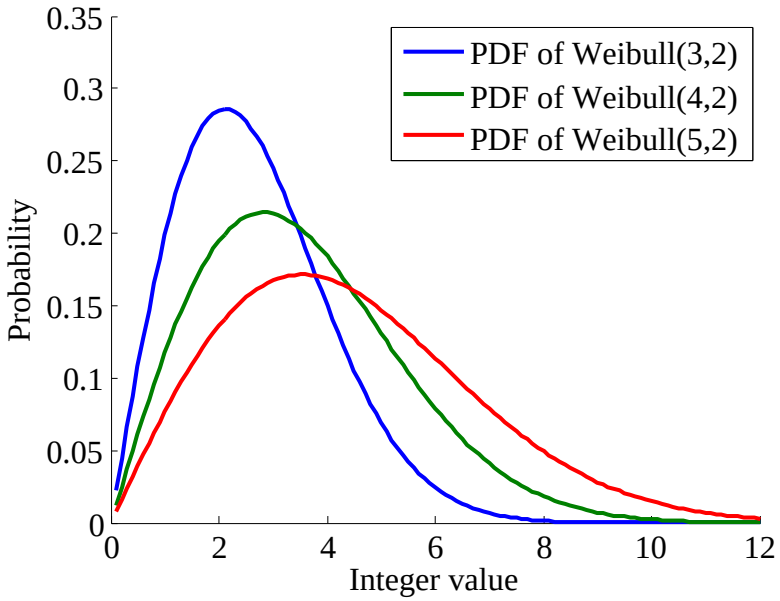

Figure 5: A probability density functions of Weibull distributions with different scale parameters A when the shape parameter B is fixed to 2.

3.2 Multivariate Probability Distributions

Multivariate probability distributions i.e. joint probability distributions are probability distributions which give the probabilities to events with multiple random variables. The multivariate probability distribution for random variables X, Y, ..., gives the probability for each of X, Y, ..., to fall in any arbitrary range of values. Multivariate probability distributions are described by joint cumulative distribution functions (CDFs) or probability density functions (PDFs) for continuous variables. Marginal distributions are a vital part when multivariate probability distributions are considered. The marginal distributions give the probabilities for every variable with no reference to the values or distributions of other random variables.

Multivariate normal distribution is the generalization of a univariate normal distribution introduced in Section 3.1.1 to higher dimensions. A random vector is multivariate normally distributed if all of the linear combinations of its components have a univariate normal distribution. The multivariate normal distribution can be used to describe any set of real-valued random variables, which are all clustered around a mean vector. The multivariate normal distribution of a k-dimensional random vector $\boldsymbol{X} = [X_1, X_2, ..., X_k]$ can be written as $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$

is a k-dimensional mean vector and $\mathbf{\Sigma}$ a $k \times k$ covariance matrix. An example of multivariate normal distribution with mean vector $\mu = [0, 0]$ and covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix} \tag{3}$$

can be found in Figure 6, where a bivariate (two-dimensional) case is presented.

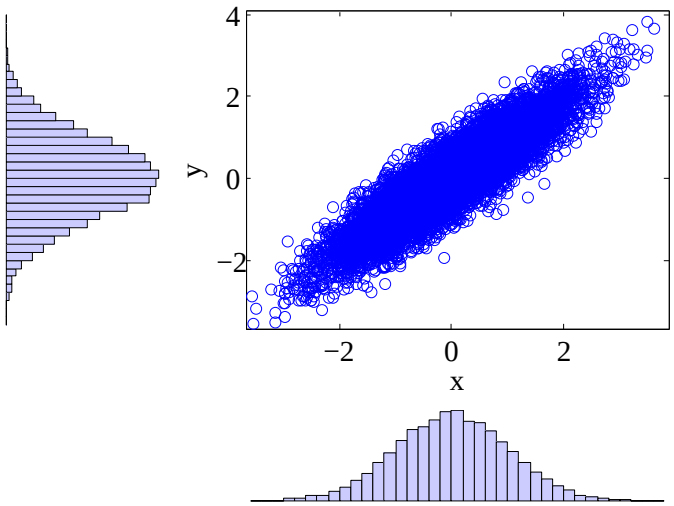

Figure 6: A bivariate normal distribution with mean vector $\mu = [0, 0]$ and covariance matrix presented in Equation (3) with the marginal distributions of both random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$.

3.3 Copula Theory

The wind speed simulation models presented in this thesis are based on copula modeling. This section introduces the concept of a copula and the copula theory required in the time series models. A copula is a multivariate probability distribution for which the marginal distribution of each variable is uniformly distributed and it describes the dependence between two or more random variables. The benefit of a copula is that the distribution of random vectors can be easily modeled and estimated by estimating the marginal distributions and copula separately.

First, the Spearman's rank-order correlation is presented, as it is a measure of dependence that is preserved through the transformations of copula modeling. Second, the basic principle of a copula is introduced and last, the Gaussian copula, which is the copula used with the time series models, is presented.

3.3.1 Spearman's Correlation Coefficient

As Pearson's correlation is not preserved through the transformations included in copula modeling, a measure of correlation that is preserved through the transformation is required. One of these preserved measures of correlation is the Spearman's rank-order correlation which is considered in this section.

Let us define the Spearman's rank-order correlation coefficient r of X and Y as the Pearson's correlation coefficient ρ between ranked variables X_{ranked} and Y_{ranked} [20]. r can be expressed as

$$r(X, Y) = \rho(X_{\text{ranked}}, Y_{\text{ranked}}). \tag{4}$$

The value of X_{ranked} for observation i is the position of the observation i in X, when X is considered in ascending order [2]. X_{ranked} gets the value 1 when the X gets its smallest value and n when X gets its highest value, when n is the sample size of X. If two observations have an equal value, they are replaced by the average of the values of the two observations. The notation $\boldsymbol{X}_{\text{ranked}}$ means that each of the variables of the random vector $\boldsymbol{X}$ is ranked version of the corresponding variable in $\boldsymbol{X}$.

3.3.2 The Basic Idea of a Copula

For a k-dimensional random vector $\boldsymbol{Y} = [Y_1, Y_2, ..., Y_k]$ with continuous cumulative distribution functions (CDFs), the CDF can be written as

$$H(y) = P[Y_1 \leq y_1, Y_2 \leq y_2, ..., Y_k \leq y_k], \tag{5}$$

where $y = [y_1, y_2, ..., y_k]$ and the right hand side of the equation represents the probability that a random variable Y_1 takes a value, which is less than or equal to y_1 et cetera. H fully describes the statistical distribution of $\boldsymbol{Y}$, but it can be very difficult to analyse directly. The copula theory provides a way to separate this analysis of H into two different parts, which notably eases the analysis.

According to Sklar's theorem, H can be written as

$$H(y) = C(u_1, u_2, ..., u_k) = C(F_1(y_1), F_2(y_2), ..., F_k(y_k)), \tag{6}$$

where C is a copula and F_i are the marginal CDFs.

The copula C defines the CDF of $\boldsymbol{U} = [U_1, U_2, ..., U_k]$, where $U_i = F_i(Y_i)$. As only continuous marginal distributions of Weibull distributed wind speeds are considered in this thesis, the margins can be considered as continuous and increasing. With these assumptions, Sklar's theorem states, that C is unique. Now we have divided H into two different parts; copula C contains the information of the dependence structure between the components of $\boldsymbol{Y}$, and the CDFs of the marginal distributions contain the information on the marginal distributions [21].

3.3.3 Gaussian Copula

One of the most commonly used copulas is the Gaussian copula. It defines the CDF of $\boldsymbol{U}$ as

$$C_\Sigma(\boldsymbol{u}) = H_\Sigma[F_N^{-1}(u_1), F_N^{-1}(u_2), ..., F_N^{-1}(u_k)], \tag{7}$$

where F_N^{-1} is the inverse CDF of the standard normal distribution and H_Σ is the joint CDF of the k-dimensional multivariate normal distribution with mean vector zero and covariance matrix $\boldsymbol{\Sigma}$ equal to the correlation matrix.

The Gaussian copula is defined by $\boldsymbol{\Sigma}$, which can be estimated by transforming $\boldsymbol{U}$ to

$$\boldsymbol{Z} = [F_N^{-1}(u_1), F_N^{-1}(u_2), ..., F_N^{-1}(u_k)], \tag{8}$$

and then calculating the correlation matrix of $\boldsymbol{Z}$. The transform presented in Equation (8) is rank-preserving as the inverse CDF if always a non-decreasing function. Therefore, we obtain $\boldsymbol{Z}_{ranked} = \boldsymbol{U}_{ranked} = \boldsymbol{Y}_{ranked}$, and accordingly, also the Spearman's rank-order correlation coefficients r are preserved through this transformation.

When simulating a random vector $\tilde{\boldsymbol{U}}$ from the Gaussian copula, a random sample $\tilde{\boldsymbol{Z}}$ drawn from a k-dimensional multivariate normal distribution. This random sample is then transformed to

$$\tilde{\boldsymbol{U}} = [F_N(\tilde{Z}_1), F_N(\tilde{Z}_2), ..., F_N(\tilde{Z}_k)]. \tag{9}$$

Then the $\tilde{\boldsymbol{U}}$ is transformed to the estimated margins $\tilde{\boldsymbol{Y}}$. As through Equation (8), also through Equation (9) the Spearman's rank-order correlation coefficients r are preserved.

Figure 7 illustrates an example of the Gaussian copula with Weibull distributed margins. The Spearman's rank-order correlation coefficients are $r_{1,2} = r_{2,1} = 0.8915$, which correspond to the Pearson's correlations (the non-diagonal components of the covarianec matrix $\boldsymbol{\Sigma}$) used in Figure 6.

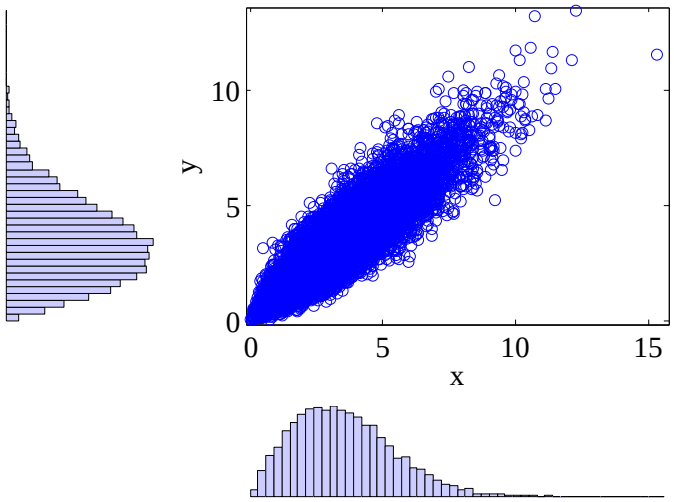

Figure 7: An example of the Gaussian copula with Weibull distributions as margins.

3.4 Autoregressive models

An autoregressive model (AR model) is a time series model that represents and describes a random time-varying process. The AR model is a special, and simpler, case of a general ARMA model. The simulation models introduced in this thesis are based on an AR model or a vector autoregressive model (VAR model). Therefore, the moving average or MA-part of the ARMA model is not considered further. Next, the basic theory of autoregressive models is presented.

3.4.1 AR Model

An AR model is a univariate time series model and an AR(p) model of order p is defined as

$$Z_t = c + \sum_{i=1}^{p} a_i Z_{t-i} + u_t, \tag{10}$$

where c is a constant, $a_1, a_2, \ldots, a_p$ are the parameters of the model and u_t is white noise. The order p defines that how many previous terms contribute to the output of the model in addition to the noise term. In an AR(4), for example, the four previous terms and the noise term contribute to the output.

3.4.2 VAR Model and Standardised VAR Model

A VAR model is a multivariate generalization of a univariate AR model presented in the previous section. It is used to depict the linear dependencies between several time series. A VAR model allows many changing variables and depending on the

order p of the VAR model, p previous terms of all of the model variables contribute to the new value of any of the variables.

A k-dimensional VAR(p) model for $\boldsymbol{Z}_t = [Z_{1,t}, Z_{2,t}, ..., Z_{k,t}]$ can be defined as

$$\boldsymbol{Z}_t = \boldsymbol{c} + \sum_{i=1}^{p} \boldsymbol{A}_i \boldsymbol{Z}_{t-i} + \boldsymbol{u}_t, \tag{11}$$

where $\boldsymbol{c}$ is a vector of intercept terms, $\boldsymbol{A}_i, \ldots, \boldsymbol{A}_p$ are the coefficient matrices of the model and $\boldsymbol{u}_t$ is white noise.

The standardised VAR model can be defined as a VAR model with $E(Z_{i,t}) = 0$ and $Var(Z_{i,t}) = 1$ for all i and t. $E(Z_{i,t}) = 0$ can be assured by ensuring the stability of the model and defining $\boldsymbol{c} = 0$.

A VAR model is stable if all the roots of its reverse characteristic polynomial, which can be written as

$$det(\boldsymbol{I}_k - \boldsymbol{A}_1 w - \boldsymbol{A}_2 w^2 ... - \boldsymbol{A}_p w^p), \tag{12}$$

where $\boldsymbol{I}_k$ is a k-dimensional identity matrix, are outside of the unit circle [17].

The VAR$_k(p)$ model can also be written in the VAR$_{kp}(1)$ state space model form as

$$\bar{\boldsymbol{Z}}_t = \bar{\boldsymbol{A}} \bar{\boldsymbol{Z}}_{t-1} + \bar{\boldsymbol{u}}_t, \tag{13}$$

which, in matrix notation can be expressed as

$$\begin{bmatrix} \boldsymbol{Z}_t \\ \boldsymbol{Z}_{t-1} \\ \boldsymbol{Z}_{t-2} \\ \vdots \\ \boldsymbol{Z}_{t-p+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots & \boldsymbol{A}_{p-1} & \boldsymbol{A}_p \\ \boldsymbol{I}_k & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_k & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{I}_k & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{Z}_{t-1} \\ \boldsymbol{Z}_{t-2} \\ \boldsymbol{Z}_{t-3} \\ \vdots \\ \boldsymbol{Z}_{t-p} \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{bmatrix}. \tag{14}$$

The autocovariances of the VAR$_k(p)$ can be obtained with

$$\Gamma_z(h) = \boldsymbol{J} \left(\sum_{i=0}^{\infty} \left(\bar{\boldsymbol{A}}^i \boldsymbol{\Sigma}_{\bar{\boldsymbol{u}}} \left(\bar{\boldsymbol{A}}^{i+h} \right)^T \right) \right) \boldsymbol{J}^T, \tag{15}$$

where $\boldsymbol{J}$ is a $k \times kp$ matrix $[\boldsymbol{J}, 0..., 0]$ and $\boldsymbol{\Sigma}_{\bar{\boldsymbol{u}}}$ is the covariance matrix of $\bar{\boldsymbol{u}}$ [17].

We defined that in standardized VAR model $Var(Z_{i,t}) = 1$. In this case, the autocorrelations of the standardized VAR$_k(p)$ model are

$$\boldsymbol{R}_z(h) = \Gamma_z(h) = \begin{bmatrix} \rho_z(1,1,h) & \rho_z(1,2,h) & \cdots & \rho_z(1,k,h) \\ \rho_z(2,1,h) & \rho_z(2,2,h) & \cdots & \rho_z(2,k,h) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_z(k,1,h) & \rho_z(k,2,h) & \cdots & \rho_z(k,k,h) \end{bmatrix}, \tag{16}$$

where ρ_z is autocorrelation and h is lag. When lag $h = 0$, in the standardized case applies that $\rho_z(i, i, 0) = Var(Z_{i,t}) = 1$, so the diagonal components of $\boldsymbol{R}_z(0)$ are all ones and the matrix is also symmetrical i.e. $\rho_z(i, j, 0) = \rho_z(j, i, 0)$ [15].

The diagonal components of $\boldsymbol{R}_z(h)$ of the process $\boldsymbol{Z}_t$ give the autocorrelation functions (ACFs) of each process and in the case of this thesis, of each location in the model. The non-diagonal components of $\boldsymbol{R}_z(h)$ give in turn the cross-correlation functions (XCFs).

The full dependence structure of a standardized VAR model is defined analytically by $\boldsymbol{R}_z(h)$, which gives the ACFs and the XCFs and the Gaussian copula, which gives the shape of the bivariate normal distribution between the locations. Figure 8 illustrates the bivariate normal distributions specified by the standardized VAR model in fully spatial, fully temporal and simultaneously spatial and temporal cases.

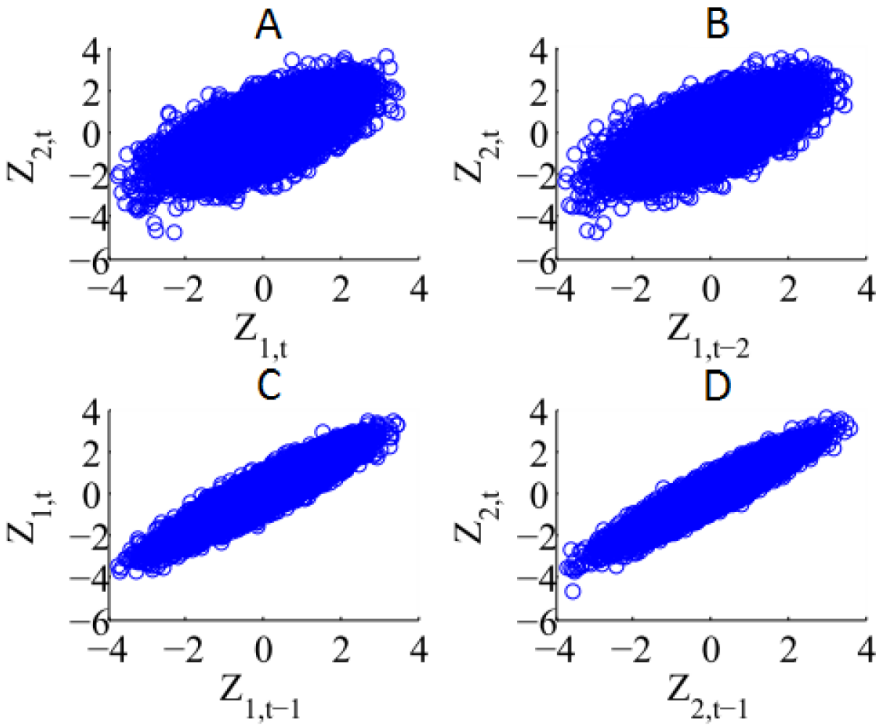

Figure 8: Four bivariate dependency structures specified by the standardized VAR model. In all four plots, both axes depict simulated normally distributed time series (in each plot subscript of $\boldsymbol{Z}$ specifies the location (1 or 2), and the moment (t, $t-1$ or $t-2$) for x- and y-axis) and the data is marked with blue circles. Plot A depicts the fully spatial case (dependency between two locations contemporaneously), plot B spatial and temporal (dependency between two locations with different moments t and $t-2$) and plots C and D both fully temporal cases (dependency between different moments t and $t-1$ in the same location. Simulations for figures were done with the high altitude data presented in Section 4.2.

3.4.3 VARX Model

The k-dimensional VAR(p) model presented in Equation (11) can be extended to a k-dimensional VARX(p), which is a VAR(p) model with exogenous variables. The VARX(p) model can be defined as

$$\boldsymbol{Z}_t = \boldsymbol{c} + \boldsymbol{b}\boldsymbol{X}_t + \sum_{i=1}^{p} \boldsymbol{A}_i \boldsymbol{Z}_{t-i} + \boldsymbol{u}_t, \tag{17}$$

where $\boldsymbol{b}$ is a $d \times k$ matrix that specifies the coefficients of the exogenous variables and $\boldsymbol{X}_t$ are the values of the d exogenous variables at time t. The exogenous variables are used in the transformed VAR model with time-dependent intercept term to create the time-dependent term, which captures the diurnal variation structures.

Next, the estimation of the VARX model parameters in presented. The same VARX model estimation is also used when estimating the transformed VAR model with time-dependent intercept term.

3.4.4 Estimation of VARX Model Parameters

In this section, the estimation of AR, VAR and VARX model parameters is presented. The estimation of the model parameters is a crucial part of the whole simulation process and thus, it is considered more in-depth. Next, the ordinary least squares (OLS) estimation of the VARX model parameters is presented. As the AR model is a univariate case of the VAR model, which is a special case of the VARX model, where parameter $\boldsymbol{b} = 0$, both of the models can be estimated with the same OLS estimation procedure as the VARX model.

The VARX(p) model presented in Equation (17) can be presented in a matrix form as

$$\begin{aligned} &\begin{bmatrix} Z_{1,t} \\ \vdots \\ Z_{k,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix} + \begin{bmatrix} b_{1,1} & \cdots & b_{1,d} \\ \vdots & \ddots & \vdots \\ b_{k,1} & \cdots & b_{k,d} \end{bmatrix} \begin{bmatrix} x_{1,t} \\ \vdots \\ x_{d,t} \end{bmatrix} + \begin{bmatrix} a^1_{1,1} & \cdots & a^1_{1,k} \\ \vdots & \ddots & \vdots \\ a^1_{k,1} & \cdots & a^1_{k,k} \end{bmatrix} \\ &\begin{bmatrix} z_{1,t-1} \\ \vdots \\ z_{k,t-1} \end{bmatrix} + \ldots + \begin{bmatrix} a^p_{1,1} & \cdots & a^p_{1,k} \\ \vdots & \ddots & \vdots \\ a^p_{k,1} & \cdots & a^p_{k,k} \end{bmatrix} \begin{bmatrix} z_{1,t-p} \\ \vdots \\ z_{k,t-p} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ \vdots \\ u_{k,t} \end{bmatrix}. \end{aligned} \tag{18}$$

where $a^1_{1,1}$ is the top left component of the $\boldsymbol{A}_1$ matrix and respectively to the other component matrices. The row i of Equation (18) can be written open as

$$\begin{aligned} &z_{i,t} = c_i + b_{i,1}x_{i,t} \ldots + b_{i,d}x_{d,t} + a^1_{i,1}z_{1,t-1} + a^1_{i,k}z_{k,t-1} \\ &+ \ldots + a^p_{i,1}z_{1,t-p} + \ldots + a^p_{i,k}z_{k,t-p} + u_{i,t}. \end{aligned} \tag{19}$$

Now, the row i presented in Equation (19) can be estimated with the OLS estimation [17]. Similarly Equation (18) can be estimated equation-by-equation for all the k

rows so all of the equations are considered. With this procedure, the estimates for the model parameters c, b and $A_1, ..., A_p$ can be obtained.

One of the benefits of the OLS estimation is the efficient estimation process of the model parameters. The predictors of the OLS estimator remain the same for all of the k equations, so the time consuming calculation of the OLS estimator has to be done only once in the whole estimation process.

3.5 Spearman's Rank-order Cross-correlation Function

This section defines the Spearman's rank-order cross-correlation function, which is the cross-correlation function between the ranked variables. As mentioned in Section 3.3.1, the Pearson's correlation is not preserved through the transformations required in the estimation and simulation of the models. However, the Spearman's rank-order correlation r is preserved as mentioned also in Section 3.3.1. Therefore, we can write that $r_Y(i,j,h) = r_Z(i,j,h) = r(i,j,h)$ for all i, j and h.

First, let us define the temporal dependence between two time series X and Y with a cross-covariance function (CCOV) as

$$CCOV(X;Y,h) = \begin{cases} \frac{1}{n}\sum_{t=1}^{n-h}(X_t - \bar{X})(Y_{t+h} - \bar{Y}) & h = 0, 1, 2, ..., \\ \frac{1}{n}\sum_{t=1}^{n+h}(Y_t - \bar{Y})(X_{t-h} - \bar{X}) & h = 0, -1, -2, ..., \end{cases} \tag{20}$$

where n is the sample size, $\bar{X}$ and $\bar{Y}$ are the means of the series, and h is the lag, for which the CCOV is calculated. A standard temporal dependence measure is the cross-correlation function (XCF), which can be written as

$$XCF(X,Y,h) = \frac{CCOV(X,Y,k)}{\sqrt{CCOV(X,X,0)}\sqrt{CCOV(Y,Y,0)}}, \tag{21}$$

where $h = 0, \pm 1, \pm 2...$ The autocorrelation function (ACF) is a special case of XCF, where the the vector X is compared with itself i.e. $X = Y$. For $h = 0$, Equation (20) estimates the covariance and Equation (21) estimates the Pearson's correlation coefficient ρ for X and Y.

In this thesis, the RXCF between two ranked variables X and Y is defined as the XCF between the corresponding ranked variables X_{ranked} and Y_{ranked}. Thus, we obtain

$$RXCF(X,Y,k) = XCF(X_{\text{ranked}}, Y_{\text{ranked}}, k). \tag{22}$$

and as the ACF is a special case of XCF, similarly RACF is a special case of RXCF where $X_{\text{ranked}} = Y_{\text{ranked}}$. So, RXCF is XCF, which is specified using r instead of ρ. With these measures, the spatial and temporal dependencies of the original measurement data and the simulation results can be compared with each other.

3.6 Transformed VAR Model

The transformed VAR model combines the VAR model presented in Section 3.4.2 and copula modeling (the transformation) presented in Section 3.3. First, the transformation necessary before the estimation of the model is presented. Second, the estimation and simulation of the transformed VAR model is discussed. Last, the distributions and the dependency structure of the time series simulated with the transformed VAR model are considered.

In the estimation process the transformed VAR model is fitted for $\boldsymbol{Z}_t$ and only one multivariate model (VAR) is fitted for the whole $\boldsymbol{Z}_t = [Z_{1,t}, Z_{2,t}, ..., Z_{k,t}]$ for all t. Therefore, $\boldsymbol{Z}_t$ for all t has to be obtained first before the estimation is possible. The measured wind speed data $\boldsymbol{Y}_t$ from each location is transformed to $\boldsymbol{Z}_t$ with transformation defined as

$$\boldsymbol{Z}_t = [F_N^{-1}(\hat{F}_1(Y_{1,t})), F_N^{-1}(\hat{F}_2(Y_{2,t})), ..., F_N^{-1}(\hat{F}_k(Y_{k,t}))]', \tag{23}$$

where $\hat{F}_i$ are the estimated marginal distributions. F_N^{-1} transforms all margins to standard normal distribution $N(0, 1)$. Therefore, $\mathrm{E}(Z_{i,t}) = 0$ and $\mathrm{Var}(Z_{i,t}) = 1$ also applies for every i and t. Then, the parameters of the standardized VAR model are estimated as described in Section 3.4.4. To clarify the transformation in Equation (23), it is illustrated in the Figure 9.

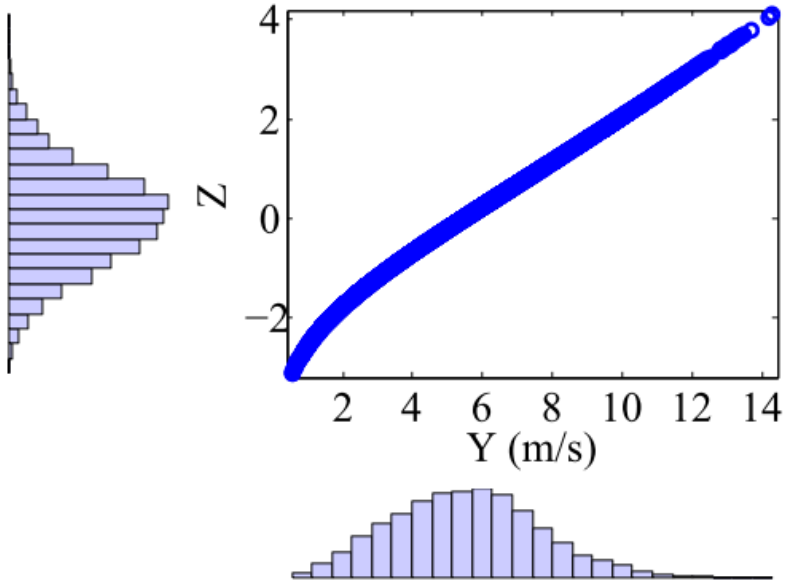

Figure 9: The transformation presented in Equation (23) from wind speeds $\boldsymbol{Y}$ (x-axis) to normally distributed $\boldsymbol{Z}$ (y-axis), which can be used in the estimation of the desired autoregressive model.

In the simulation phase, the standardized VAR model [17] is used and $\tilde{\boldsymbol{Z}}_t$ is simulated from the standardized $\mathrm{VAR}_k(p)$ model. As presented in Section 3.4.2, $\mathrm{E}(Z_{i,t}) = 0$

and $\mathrm{Var}(Z_{i,t}) = 1$ applies for the standardized model and the error terms u are normally distributed. Therefore, the simulation result $\tilde{Z}_{i,t}$ follows $N(0,1)$ for every i and t. [15]

Next, the simulated time series are transformed back to the wind speed domain with following transformation

$$\tilde{\boldsymbol{Y}}_t = [\hat{F}_1^{-1}(F_N(\tilde{Z}_{1,t})), \hat{F}_2^{-1}(F_N(\tilde{Z}_{2,t})), ..., \hat{F}_k^{-1}(F_N(\tilde{Z}_{k,t}))]', \tag{24}$$

The estimated margins $\hat{F}_i$ are the same for all t as $\tilde{Z}_{i,t} \sim N(0,1)$. The transformed time series $\tilde{\boldsymbol{Y}}_t$ follow the estimated margins $\hat{\boldsymbol{F}}$. Next, the distributions of the transformed VAR model is considered.

If two arbitrary simulated random variables of the standardized VAR model are chosen, the distribution of those variables follows a bivariate normal distribution $N_2(\boldsymbol{0}, \boldsymbol{C})$ [15], where correlation matrix $\boldsymbol{C}$ is

$$\boldsymbol{C} = \begin{bmatrix} 1 & \rho_z(i,j,h) \\ \rho_z(i,j,h) & 1 \end{bmatrix}. \tag{25}$$

After $\tilde{\boldsymbol{Z}}_t$ has been transformed with Equation (24) to $\tilde{\boldsymbol{Y}}_t$, any two random variables i.e. simulated wind speed values from $\tilde{\boldsymbol{Y}}_t$ are correlated with each other (analytically determined RACFs and RXCFs) and follow their estimated location specific marginal distributions. However, these two measures are not enough to specify the full dependency structure between the random variables. This dependency structure is defined as a bivariate Gaussian copula, presented in Equation (7), specified by correlation matrix $\boldsymbol{C}$. Figure 10 illustrates the bivariate normal distributions of the simulated wind speeds specified by the transformed VAR model in fully spatial, fully temporal and simultaneously spatial and temporal cases.

The Spearman's rank-order correlation was introduced in Section 3.3.1. In case of the Gaussian copula, the Spearman's rank-order correlation coefficient r can be determined from the Pearson's correlation coefficient ρ with equation

$$r = \frac{6}{\pi} \arcsin \frac{\rho}{2} \tag{26}$$

Now, the random variables in $\tilde{\boldsymbol{Y}}_t$, simulated with the transformed VAR model, have the estimated marginal distributions, correlations defined by the Spearman's rank-order correlation $r(i,j,h)$ and dependency structure specified by the Gaussian copula with a certain $\boldsymbol{C}$.

The dependency structure between the random variables is specified analytically by the Gaussian copula, but there are also numerous other copulas, which could specify the dependency structure between the random variables. The theory of copula modeling states that with the same Spearman's rank-order correlation and

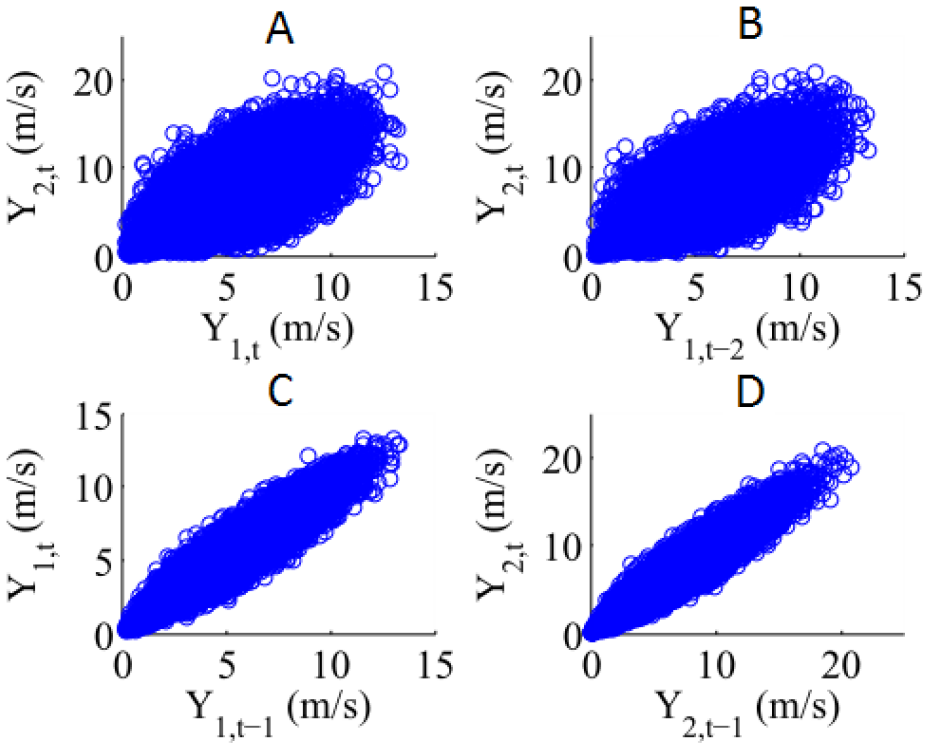

Figure 10: Four bivariate dependency structures as in Figure 8, but now in wind speeds and specified by the transformed VAR model. In all four plots, both axes depict simulated wind speeds (m/s) (in each plot subscript of $\boldsymbol{Y}$ specifies the location (1 or 2), and the moment (t, $t-1$ or $t-2$) for x- and y-axis) and the data is marked with blue circles. Plot A depicts the fully spatial case (dependency between two locations contemporaneously), plot B spatial and temporal (dependency between two locations with different moments t and $t-2$) and plots C and D both fully temporal cases (dependency between different moments t and $t-1$ in the same location. Simulations for the figures were done with the high altitude data presented in Section 4.2.

identical marginal distributions, the dependency structure can still be specified by various different copulas. [21]

In this thesis, it is assumed that the Gaussian copula is the right copula to define the dependency structure in the cases and data presented. Also, if Gaussian copula would not specify the dependency structure, the ARC and VAR models could not be used to simulate the wind speeds as the models require normally distributed data as an input. In addition, in Chapter 8 t-copula and copulas from Archimedean family are shortly discussed.

3.7 Transformed ARC Model

This section introduces the basic structure of the transformed ARC model. Section 5.3 presents the transformed ARC model as part of the whole simulation model

including the consideration of the diurnal variations. Also, the whole estimation and simulation process with the transformed ARC model is presented step-by-step in Section 5.3. This section, in turn, presents the estimation and simulation of the transformed ARC model itself, without other parts of the simulation model involved.

First, the wind speed measurement data $\boldsymbol{Y}_t$ is transformed to $\boldsymbol{Z}_t$ with Equation (23) as done with the transformed VAR model.

The transformed ARC model consists of univariate AR(p) models and the correlation matrix $\boldsymbol{C}$. In the estimation of the model a univariate AR(p) model is fitted for each k wind power generation locations ($\boldsymbol{Z}_i$, with i noting the location) separately and then the k independent AR(p) models are linked to each other with the correlation matrix $\boldsymbol{C}$. [11]

The suitable order for the AR models used with the data presented in Chapter 4. was five. The determination of the order of the model was done by observing the autocorrelations of the residuals with lag $h \neq 0$. The order should be increased until all autocorrelations from the residuals are removed and this was achieved with the AR(5) model.

The final step of the estimation process is the determination of $\boldsymbol{C}$ from the normally distributed data $\boldsymbol{Z}_t$.

In simulation, data is first simulated from the k estimated AR(5) models. Thus, simulated time series for each location is obtained, though the time series still lack the correlations between the locations. The spatial dependency to the time series is achieved by calculating the Cholesky decomposition of $\boldsymbol{C}$ and then multiplying the time series with the Cholesky decomposition as done in [11]. After the addition of the diurnal variations, considered more closely in Section 5.3.1, the obtained time series $\tilde{\boldsymbol{Z}}_t$ are transformed back to wind speeds using Equation (24). Next, the Cholesky decomposition is presented.

When simulating with the transformed ARC model, the Cholesky decomposition is used to create the correlation between the simulated locations. The Cholesky decomposition is a decomposition of a positive-definite matrix into a product of a lower triangular matrix and its conjugate transpose. The Cholesky decomposition can be written as

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T, \tag{27}$$

where $\boldsymbol{A}$ is a positive-definite matrix, $\boldsymbol{L}$ is a lower triangular matrix with real and positive diagonal components and $\boldsymbol{L}^T$ is the conjugate transpose of the matrix $\boldsymbol{L}$.

In the case of the transformed ARC model, $\boldsymbol{A}$ in Equation (27) is the correlation matrix $\boldsymbol{C}$. The correlated time series are obtained by calculating the following equation

$$\boldsymbol{Z}_{\text{corr}} = \boldsymbol{L}\boldsymbol{Z}_{\text{sim}} \tag{28}$$

where $\boldsymbol{Z}_{\text{corr}}$ is the matrix containing the time series with correlation between locations and $\boldsymbol{Z}_{\text{sim}}$ the matrix of uncorrelated time series obtained by using Monte Carlo

simulations to the univariate AR(5) models. [11]

Also, if each sample of $\boldsymbol{Z}_{\text{sim}}$ follows a Normal distribution, also $\boldsymbol{Z}_{\text{corr}}$ follows a Normal distribution as presented in [11]. Next, the distributions of the transformed ARC model in considered.

In case of transformed ARC model, the autocovariances of the $AR(p)$ model can be obtained analytically with a univariate version of the Equation (15). The autocorrelations are determined for each $AR(p)$ model separately and from these the rank-order autocorrelation functions (RACFs) can be obtained for every lag h.

The rank-order cross-correlation function (RXCF) with $h = 0$ can be determined from the parameters of the correlation matrix $\boldsymbol{C}$. The RXCFs when $h \neq 0$ are not defined analytically for transformed ARC model, but they can be determined numerically with e.g. Monte Carlo simulations, as which is the case in this thesis.

The dependency structure of the transformed ARC model is assumed to be defined by a bivariate Gaussian copula as it was with the transformed VAR model. Figure 11 illustrates the bivariate normal distribution of the simulated wind speeds specified by the transformed ARC model in fully spatial, fully temporal and simultaneously spatial and temporal cases. If Figure 11 is assessed graphically, it is justified to assume that the dependency structure is defined by a bivariate Gaussian copula.

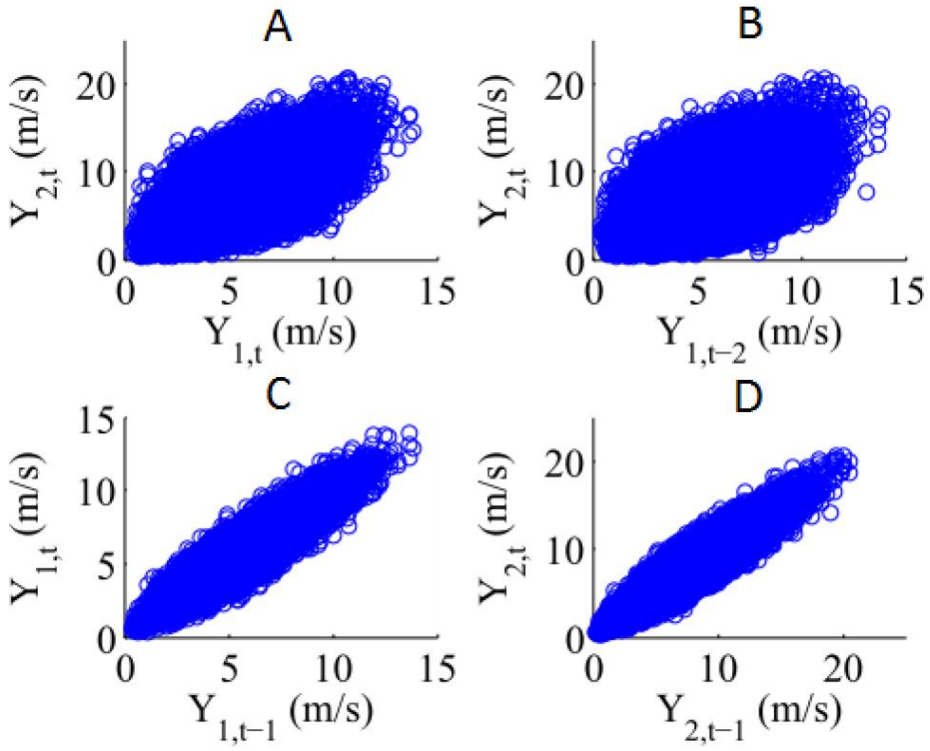

Figure 11: Four bivariate dependency structures as in Figure 8, but now in wind speeds and specified by the transformed ARC model. In all four plots, both axes depict simulated wind speeds (m/s) (in each plot subscript of $\boldsymbol{Y}$ specifies the location (1 or 2), and the moment (t, $t-1$ or $t-2$) for x- and y-axis)) and the data is marked with blue circles. Plot A depicts the fully spatial case (dependency between two locations contemporaneously), plot B spatial and temporal (dependency between two locations with different moments t and $t-2$) and plots C and D both fully temporal cases (dependency between different moments t and $t-1$ in the same location. Simulations for the figures were done with the high altitude data presented in Section 4.2.

3.8 Monte Carlo Simulations

This section introduces the Monte Carlo simulation method, which is used with both models presented in this thesis. Monte Carlo method is a numerical problem solving technique and the basic principle behind it is that the behaviour of a statistic in random samples can be assessed by drawing multiple random samples and observing the behaviour of the statistic in those samples. In Monte Carlo simulations sets of random numbers called pseudo-populations, which emulate the samples in real world population, are created. These artificially generated pseudo-populations (that resemble the real samples) are used to carry out several simulation runs of the process considered and the behaviour of the process with different samples is observed and stored. [22]

The basic Monte Carlo method can be written as follows:

1. The pseudo-population is specified in a way that it can be used to generate samples. This can be done with a computer algorithm.
2. A sample is drawn from the pseudo-population. The sample from the pseudo-population should resemble the behaviour of the actual real world population.
3. Carry out simulation runs of the process considered and store the results.
4. Repeat steps 2. and 3. t times, where t is the number of simulations.

In this thesis, all the simulations used with the transformed ARC and VAR models are done with Monte Carlo method. Monte Carlo simulations are used to generate samples from a normal probability distribution for the models as both of the models require normally distributed data as an input for the simulations.

3.9 Summary of the Theory of the Time Series Models

This chapter presents the theory of time series models needed for the simulation models applied to wind speeds. The normal distribution was presented, as the input data for the autoregressive models should be normally distributed. Wind speeds and the measurement data are Weibull distributed and therefore, it is necessary to introduce the Weibull distribution. Also, the multivariate normal distribution was presented as the dependency structure specified by the Gaussian copula is a multinormally distributed.

The required copula theory was introduced, as copula transformation is used to convert the Weibull distributed wind speed measurement data to normally distributed. Gaussian copula was presented separately, as it is the specific copula that defines the dependency structure between different measurement locations or between different time steps. Also, ranked correlation measures were introduced as rank-correlation is preserved through copula transformations.

Third section of Chapter 3 considered autoregressive models. The section began with the simple univariate AR model and then presented relevant theory concerning VAR and VARX models and the estimation of the model parameters. Also, ranked version of the cross-correlation function (XCF) was introduced.

After the fundamental theory of different parts required in the wind speed models, Chapter 3 introduced the transformed VAR and ARC models, which are the frames for the simulation models applied to wind speeds. At the end of Chapter 3, also the Monte Carlo simulations were shortly introduced, as it is the method used in the wind speed simulations.

4 Data

For the models presented in this thesis, input data for the estimation of the models is required. The data used in this thesis can be divided into three categories, low and high altitude wind speed data and aggregate wind power generation and installed capacity data from Finland. Next, the different data are introduced more precisely.

4.1 Low Altitude Wind Speed Data

The low altitude wind speed data is obtained from The Finnish Meteorological Institute (FMI). The data consists of 19 measurement locations from Finland with the average altitude of 15 meters above the surrounding ground level. The data is measured between July 2008 and July 2011 from all 19 locations. The time resolution of the data is hourly, so the number of observations is $T = 23735$. The low altitude data is used in the verification of the transformed VAR model with time-dependent intercept term and transformed ARC model when modeling existing locations and also with the transformed ARC model when adding new locations to the model.

The low altitude measurement data is measured with physical measurement devices that have a lower measurement limit, which means that when the wind speeds are below the lower measurement limit, the recorded wind speed value in the data is zero. Each location has a specific lower measurement limit value which is 0.6 or 1 m/s depending on the measurement device in the location. This has to be considered when fitting the marginal distributions for the measurement locations.

A random variable from Weibull distribution has to be larger than zero as Weibull distribution is a positive probability distribution and in addition, values near zero should be very rare. Therefore, the zero values caused by the lower measurement limit of the measurement devices has to be taken into account. This problem is considered with the usage of left-censoring of the data when estimating the Weibull parameters for the marginal distributions. Left-censoring transforms the measurements recorded as zero to values which are somewhere between zero and lower measurement limit. The left-censoring is done by using the maximum likelihood estimation algorithm of the Weibull parameters provided in [23].

4.2 High Altitude Wind Speed Data

The used high altitude measurement data is from two locations, Hyytiälä and Puijo. The data from Hyytiälä was obtained from The University of Helsinki and the data from Puijo from The Finnish Meteorological Institute (FMI). The data from Hyytiälä is measured 74 meters above the surrounding ground level and the data from Puijo 75 meters. Puijo is also located on a 150 meters high hill which makes it higher compared with the sea level. Therefore, as wind speeds are higher in higher altitudes, the wind speeds in Puijo are higher than in Hyytiälä. The data is

measured between January 2007 and July 2009 from both locations and therefore, the number of observations is $T = 22150$. The high altitude measurement data is used in the comparison of the dependency structures between locations and in the verification of the models when modeling existing locations.

As mentioned earlier, Figure 10 presents the bivariate normal distributions of wind speed simulations specified by the transformed VAR model and Figure 11 presents the bivariate normal distributions of wind speed simulations specified by the transformed ARC model. Simulations for Figures 10 and 11 were done with the high altitude data from locations Hyytiälä (location 1) and Puijo (location 2). Figure 12 illustrates the bivariate normal distributions of measurement data from Puijo and Hyytiälä in fully spatial, fully temporal and simultaneously spatial and temporal cases. By looking at these two figures, it can be observed that the Gaussian copula does define the dependency structure between the locations.

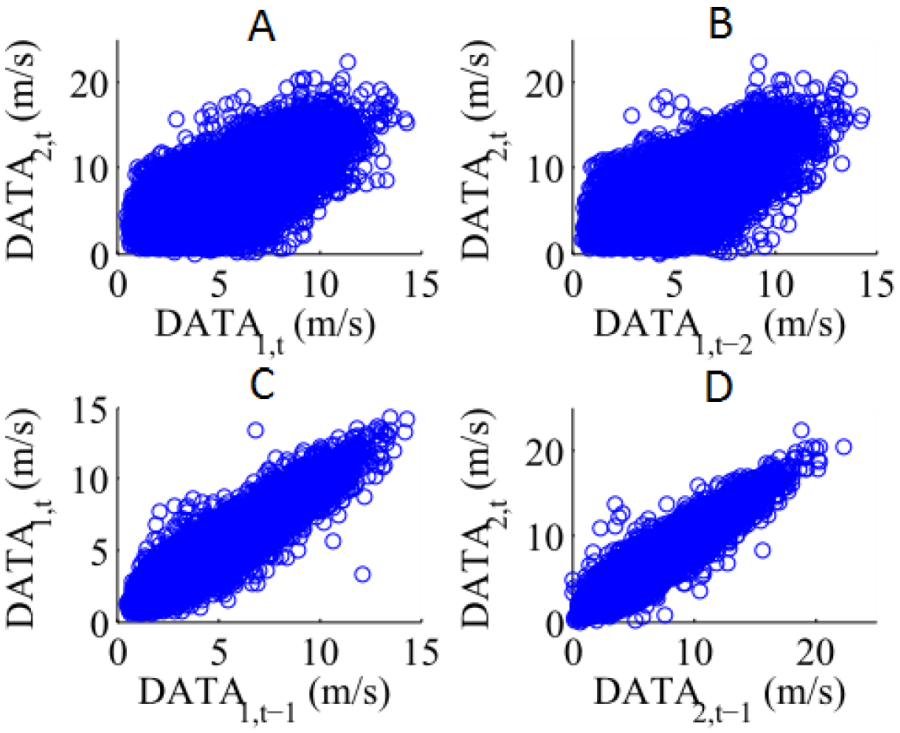

Figure 12: Four bivariate dependency structures from high altitude measurement data from Hyytiälä (1) and Puijo (2). In all four plots, both axes depict simulated wind speeds (m/s) (in each plot subscript of ***DATA*** the location (1 or 2), and the moment (t, $t-1$ or $t-2$) for x- and y-axis) and the data is marked with blue circles. Plot A depicts the fully spatial case (dependency between two locations contemporaneously), plot B spatial and temporal (dependency between two locations with different moments t and $t-2$) and plots C and D both fully temporal cases (dependency between different moments t and $t-1$ in the same location.

4.3 Aggregate Wind Power Generation and Capacity Data from Finland

Power generation data from single turbines is hard to obtain, but aggregate wind power generation data from Finland between 2008 and 2012 was provided by Energiateollisuus (ET). ET provided also the capacity data of installed wind generation capacity in Finland from 1991 to end of 2012. The aggregated generation and capacity data are both available for everyone free of charge but the data has to be requested from ET. The aggregated generation data is used in the assessment of the cases presented in Chapter 7.

5 Simulation Models Applied to Wind Speeds

This chapter presents the two wind speed simulation models introduced in this thesis, the transformed ARC model and the transformed VAR model with time-dependent intercept term. The theoretical background of the different components of these models was introduced in Chapter 3. In this chapter, the components are put together to form the full simulation models.

This chapter is divided into three different parts. In the first section of this chapter, the fitting of the marginal distributions is presented. Then, in the second, the transformed VAR model with time-dependent intercept term is introduced. The whole process from the estimation of the marginal distributions to the final conversion of the simulated time series to wind speeds is introduced step-by-step.

Other important matters that require special attention concerning the transformed VAR model with time-dependent intercept term are the implementation of the diurnal variation and the normality of the simulation results when the diurnal variations are considered. Both of these matters are discussed in their own subsections. Also, the problems with implementation of new non-measured locations to the transformed VAR model with time-dependent intercept term is discussed.

The last section of this chapter considers the transformed ARC model. In this section the whole process of the simulation and estimation of the transformed ARC model is presented step-by-step from the marginal distributions to the final simulation results. As with the transformed VAR model with time-dependent intercept term, the implementation of the day structures and the normality of the results after the implementation are discussed in their own subsections.

5.1 Fitting of the Marginal Distributions

When analyzing wind speeds with either one of the models, the analysis starts with the estimation of the marginal distributions. Wind speeds are commonly considered as Weibull distributed, which is also assumed in this thesis. The Weibull distributions are estimated in Matlab by using the method of maximum likelihood.

There are various approaches how to consider the marginal distributions. Next, two different approaches are presented. [14] and [2] propose that a different distribution should be fitted for each hour of the day. The data is divided into groups representing each hour and a distribution if fitted for each group. With this approach, the diurnal variations of the measurement data can be analyzed with the marginal distributions because each distribution now has a different expected value and standard deviation. 24 groups (one for each hour in a day) have been used in [2].

In the data used in this thesis, there are hourly and monthly variations in all of the measurement locations. Therefore, $12 \times 24 = 288$ distributions should be fitted for each location for the full analysis of the diurnal variation structures. This approach

was not considered feasible with the data used because it would leave too few hours of measurements for each of the 288 groups to conduct accurate analysis. Thus, a different approach, in which one Weibull distribution is fitted for each location, was chosen as in [3], [11], [24] and [25]. With this approach, the marginal distribution F_i describes the distribution of wind speeds in one location for all t.

With the chosen approach, the monthly changing diurnal variations are not considered in the marginal distributions, and therefore have to be taken into account separately. Section 5.2.1 and 5.3.1 consider the implementation of the changing diurnal variations.

5.2 Transformed VAR Model with Time-dependent Intercept Term

Transformed VAR model with time-dependent intercept term is the first of the two models presented in this thesis. It is a multivariate model which includes both the temporal and spatial dependency and, in addition, the expected values $\mathrm{E}(\boldsymbol{Y}_t)$ change according to the month and hour considered i.e. the monthly diurnal variations are also considered in the model.

The whole process is complex and has many steps including the fitting of the margins, estimation and simulation. Therefore, the whole process is presented in the Figure 13 to illustrate the processes involved with the model.

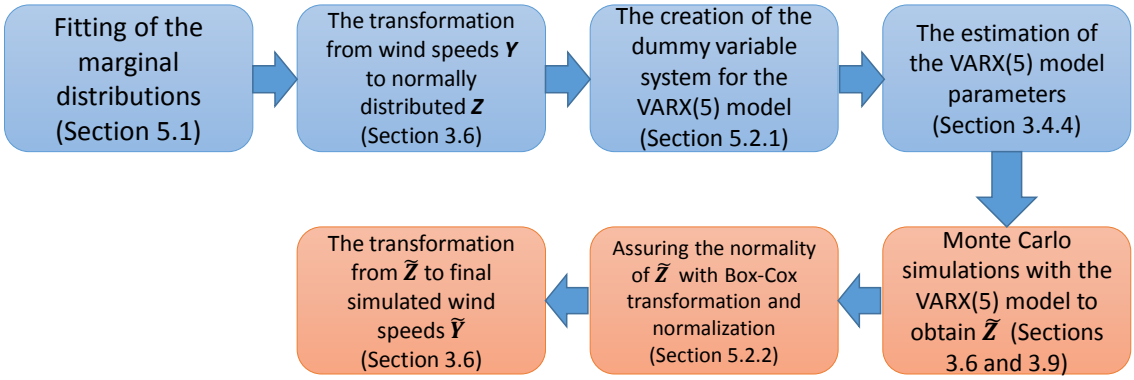

Figure 13: The whole process of the transformed VAR model with time-dependent intercept term depicted with blocks describing each stage of the process. Blue blocks are part of the estimation process and orange blocks part of the simulation process.

5.2.1 Implementing the Monthly Diurnal Variations

The diurnal variations are implemented to the model through the time-dependent intercept term. The Transformed VAR model with time-dependent intercept term

can be written as

$$\boldsymbol{Z}_t = \boldsymbol{c}_t + \sum_{i=1}^{p} \boldsymbol{A}_i \boldsymbol{Z}_{t-i} + \boldsymbol{u}_t, \tag{29}$$

where $\boldsymbol{c}_t$ is the intercept term. If the intercept term is written open as

$$\boldsymbol{c}_t = \boldsymbol{c} + \boldsymbol{b}\boldsymbol{X}_t, \tag{30}$$

the model in Equation (29) is identical with the VARX model in Equation (17). Therefore, it can be also estimated with the VARX model estimation as presented in Section 3.4.4.

The monthly diurnal variations are analyzed by implementing them in the model as a dummy variable system [17], [18]. Consider a dummy vector

$$\boldsymbol{D} = [D_1, D_2, ..., D_{288}], \tag{31}$$

where D_i is a dummy variable that gets a value 1 when the considered hour belongs to the month and hour of the day that D_i represents. To clarify the usage of the dummy variables, an example is presented. Consider D_1 that gets the value 1 between midnight and 1 am in January and 0 otherwise. This results in $\boldsymbol{c}_t = \boldsymbol{c}_0$. D_{25}, in turn, gets the value 1 between midnight and 1 am in February and 0 otherwise. In this case, the result is $\boldsymbol{c}_t = \boldsymbol{c}_0 + \boldsymbol{b}_{24}$, and so on.

As the dummy system presented in [17] specifies each hour in the data and the intercept term $\boldsymbol{c}$ already exists in the model in Equation (17), one dummy variable has to be removed. [18] In the case of this model, the first dummy variable is removed. Therefore, the exogenous variables $\boldsymbol{X}$ used in the model are

$$\boldsymbol{X} = [D_2, D_3, ..., D_{287}]. \tag{32}$$

The presented inclusion of exogenous variables to the model allows the expected values $E(\boldsymbol{Z}_t)$ to be time-varying according to the monthly diurnal variations.

In [17], it is noted that the VAR model with time-dependent intercept term is not stationary because $E(\boldsymbol{Z}_t)$ gets different values when t changes. Although, the model is not stationary, it is still stable, as the VAR part of the process, according to [9], and the dummy variable system are both stable.

5.2.2 Assuring the Correct Marginal Distributions in Simulation

As the expected value $E(\boldsymbol{Z}_t)$ gets different values when t changes, the simulated time series $\tilde{\boldsymbol{Z}}_i$ from location i is not necessarily Gaussian. Though, for the transformation back to the wind speed domain, the normality of the time series is required. Therefore, the normality of $\tilde{\boldsymbol{Z}}_i$ for all i has to be ensured. This can be carried out

with the Box-Cox transformation [26], which is a rank-preserving transformation, so the RACFs and RXCFs remain the same. The transformation can be written as

$$data_\lambda = \frac{data^\lambda - 1}{\lambda}, \tag{33}$$

where *data* is the input time series data to be transformed to normal distribution and $\lambda > 0$. This transformation was implemented for all i locations and performed with Matlab function *boxcox*. As seen from the Equation (33), if $\lambda = 1$, there transformation does not change the shape of the input data. So, if λ is close to 1, the transformation has a minimal effect on the data. As the Box-Cox transformations were performed for the data, the average $\lambda \approx 1$, so the data was already very close to normal distribution, for all i. After the Box-Cox transformations, the time series were also normalized to have mean 0 and variance 1. With these procedures, it is ensured, that when the time series $\tilde{\boldsymbol{Z}}_i$ are transformed to the final simulated wind speeds $\tilde{\boldsymbol{Y}}_i$, they follow the correct marginal distributions.

Figure 14 illustrates the bivariate normal distributions of the simulated wind speed data specified by the transformed ARC model with diurnal variations in fully spatial, fully temporal and simultaneously spatial and temporal cases. After the implementation of the day structures to the model, it is not theoretically justified to state that the dependency structure is still defined by the Gaussian copula. However, Figure 14 shows that the dependency structure has not changed significantly compared with Figure 10. Therefore, it can be assumed, that the dependency structure is specified by the Gaussian copula also in the case of the transformed VAR model with time-dependent intercept term. The dependency structure of the bivariate normal distribution of the simulated time series before the transformation to wind speeds is presented in Appendix A.

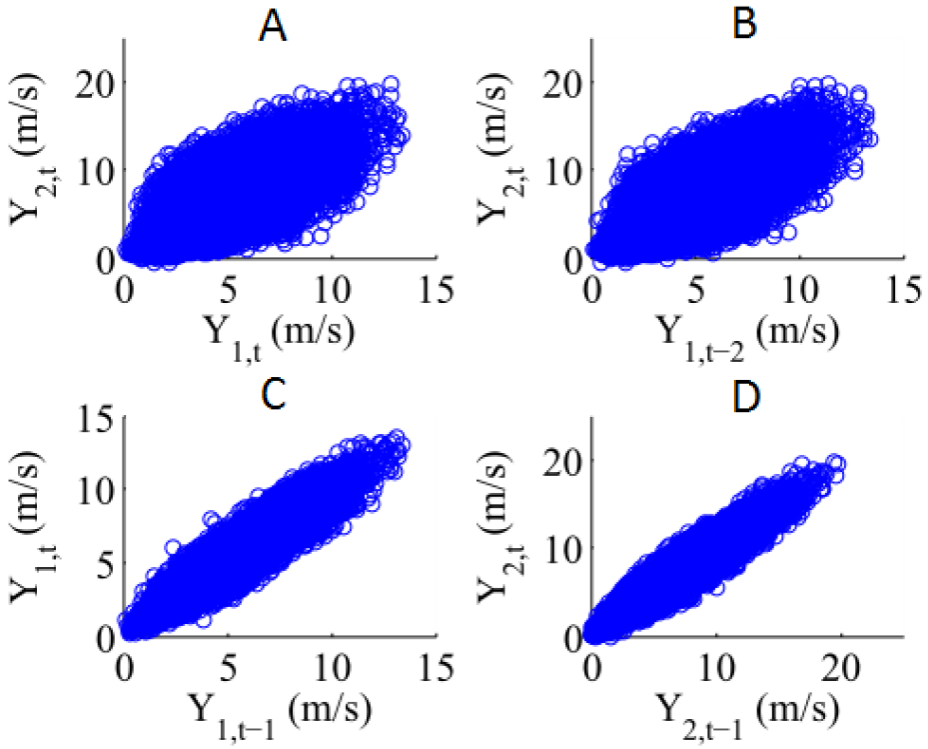

Figure 14: Four bivariate dependency structures specified by the transformed VAR model with time-dependent intercept term. In all four plots, both axes depict simulated wind speeds (m/s) (in each plot subscript of $\boldsymbol{Y}$ specifies the location (1 or 2), and the moment (t, $t-1$ or $t-2$) for x- and y-axis) and the data is marked with blue circles. Plot A depicts the fully spatial case (dependency between two locations contemporaneously), plot B spatial and temporal (dependency between two locations with different moments t and $t-2$) and plots C and D both fully temporal cases (dependency between different moments t and $t-1$ in the same location. Simulations for figures were done with the high altitude data presented in Section 4.2.

5.2.3 Problems with New Locations

When adding new wind power generation locations to the transformed VAR model with time-dependent intercept term, major problems are encountered. This section shows that it is not feasible to use the transformed VAR model with time-dependent intercept term when new locations are added.

The autocovariances Γ_z presented in Equation (15) of the $\text{VAR}_k(p)$ can be simplified for VAR(1) as

$$\Gamma(h) = \sum_{i=0}^{\infty} \left(\boldsymbol{A}_1^i \Sigma_{\bar{u}} \left(\boldsymbol{A}_1^{i+h} \right)^T \right). \tag{34}$$

The parameters of the matrix $\boldsymbol{A}_1$ can be written as a_{ij}, where i is the row and j is the column of the element. The geographical distance between two locations i and j is denoted by x_{ij}.

If parameters a_{ij} could link the correlation structure between two location i and j to the geographical distance x_{ij}, it would require that the change in a_{ij} would have an effect only to the corresponding parameter of Γ. However, the previous requirement is not fulfilled, as seen in Equation (34). If any a_{ij} is changed, it can affect to all of the parameters in Γ through the multiplication of the matrices. As a direct consequence, the distances x_{ij} cannot be linked directly to parameters a_{ij}.

It is possible that the distances x_{ij} could be linked to the parameters of Γ instead of a_{ij} through Yule-Walker equations. However, it would be a complex procedure, which would require the estimation of all $\boldsymbol{A}_1$ and $\boldsymbol{\Sigma}_{\bar{u}}$ parameters every time when a new location is added to the model. Therefore, this was not considered as a feasible approach and it is not considered further in this thesis, except a short discussion in Chapter 8.

5.3 The Transformed ARC Model with Diurnal Variations

This section presents the second of the two models presented in this thesis. The transformed ARC model is already introduced and now the whole process is presented from the fitting of the margins to the transformation of the simulation results to wind speeds. The model is complex and consists of several steps which are presented in Figure 15 to illustrate the usage of the model. The model, already introduced in Section 3.6, already contains the spatial and temporal dependencies and the dependency structure described by the Gaussian copula. Also the diurnal variations are considered in the model as presented in the next section.

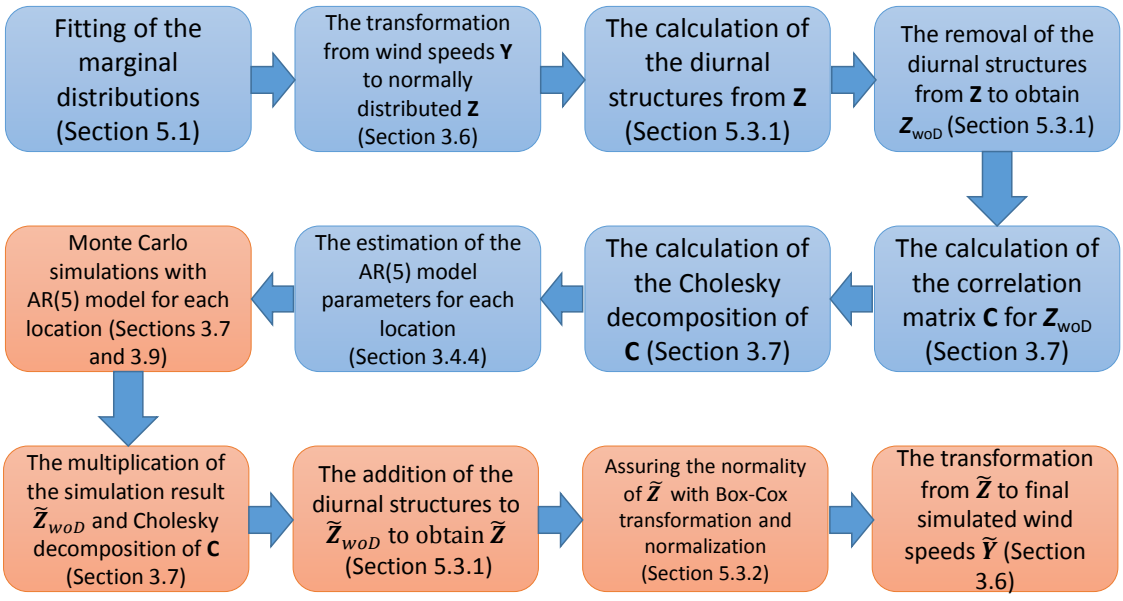

Figure 15: The whole process of the transformed ARC model depicted with blocks describing each stage of the process. Blue blocks are part of the estimation process and orange blocks part of the simulation process.

It was shown that the transformed VAR model with time-dependent intercept term was not suitable for simulations with new locations. However, this is not the case with the transformed ARC model. It can be used when adding new locations to the model because the distances x_{ij} between locations can be linked to the parameters of the correlation matrix $\boldsymbol{C}$. This is considered in section 5.3.3.

5.3.1 Implementing the Monthly Diurnal Variations

This section presents how the diurnal variations are considered in the transformed ARC model. The diurnal variations are first removed from the estimation data before the estimation of the AR model parameters and added back after simulation [13]. First, the monthly changing diurnal structure is calculated for each hour of each month resulting in $12 \times 24 = 288$ averages for each location. Next, the calculated averages are subtracted from the $\boldsymbol{Z}_t$ and thus, $\boldsymbol{Z}_{t,\text{WoD}}$ is obtained. $\boldsymbol{Z}_{t,\text{WoD}}$ is used in the estimation of the AR(5) models for each location and also the correlation matrix $\boldsymbol{C}$ is calculated for $\boldsymbol{Z}_{t,\text{WoD}}$.

The day structures are added back to the simulated data matrix after the multiplication of the time series with the Cholesky decomposition of $\boldsymbol{C}$. Thus, time series $\tilde{\boldsymbol{Z}}_i$, consisting of the temporal and spatial dependence structure and the diurnal variations is obtained. Then, $\tilde{\boldsymbol{Z}}_i$ is transformed to the wind speed data time series $\tilde{\boldsymbol{Y}}_i$. [25]

5.3.2 Assuring the Correct Marginal Distributions in Simulation

The removal of the diurnal structure in the estimation of the model and adding it back after the simulation can have an effect on the normality of the simulated time series. Also, the multiplication of the time series with the Cholesky decomposition does not necessarily preserve the normality of the simulated time series. The normality can be ensured with the Box-Cox transformation, already introduced in Section 5.2.2, and normalization of the time series following the same procedure as with the transformed VAR model with time-dependent intercept term as presented in Section 5.2.2.

As with the transformed VAR model with time-dependent intercept term, the simulated data was already very close to normal distribution, for all i. Therefore, the necessity of the Box-Cox transformation is debatable, but the simulated time series required the normalization to have mean 0 and variance 1. Thus, it could be ensured that when the time series $\tilde{\boldsymbol{Z}}_i$ are transformed to the final wind speeds $\tilde{\boldsymbol{Y}}_i$, they follow the correct marginal distributions.

Figure 16 illustrates the bivariate normal distributions of the simulated wind speed data specified by the transformed ARC model with diurnal variations in fully spatial, fully temporal and simultaneously spatial and temporal cases. It can be seen from the Figure 16 that the dependency structure has not changed significantly, based on

visual observations, compared with Figure 11. Therefore, it can be assumed, that the dependency structure is still specified by the Gaussian copula also in the case of the transformed ARC model with diurnal diurnal variations. The dependency structure of the bivariate normal distribution of the simulated time series before the transformation to wind speeds is presented in Appendix A.

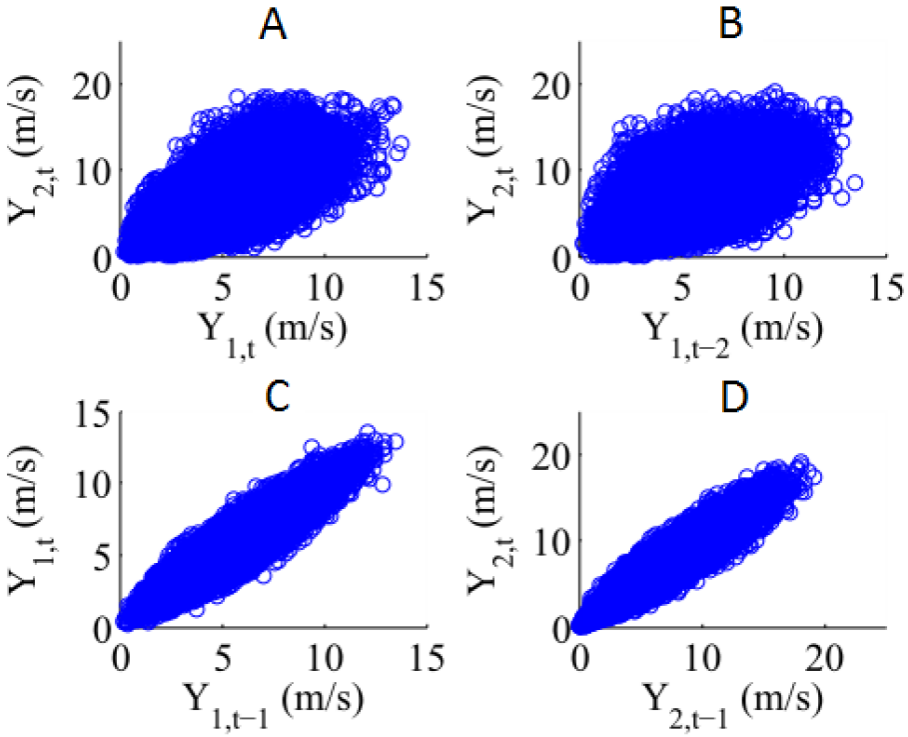

Figure 16: Four bivariate dependency structures specified by the transformed ARC model with diurnal variations. In all four plots, both axes depict simulated wind speeds (in each plot subscript of $\boldsymbol{Y}$ specifies the location (1 or 2), and the moment (t, $t-1$ or $t-2$) for x- and y-axis) (m/s) and the data is marked with blue circles. Plot A depicts the fully spatial case (dependency between two locations contemporaneously), plot B spatial and temporal (dependency between two locations with different moments t and $t-2$) and plots C and D both fully temporal cases (dependency between different moments t and $t-1$ in the same location. Simulations for figures were done with the high altitude data presented in Section 4.2.

5.3.3 The Addition of New Locations to the Transformed ARC Model

The transformed ARC model is suitable for the addition of new locations to the model. The parameters $\rho_{i,j}$ of the correlation matrix $\boldsymbol{C}$ can be linked to the distance x_{ij} between the locations i and j. Unlike with the transformed VAR model with time-dependent intercept term, the distance x_{ij} linked to $\rho_{i,j}$ has an effect only to locations i and j, the parameters linking other locations remain the same.

When adding new non-measured locations to the model, the required $\rho_{i,j}$ linking

new locations to existing ones can be obtained from an exponential curve fitted to the existing parameters of $\boldsymbol{C}$. In this thesis, the fitted curve was obtained with the fitting toolbox of Matlab and the equation of the curve can be written as

$$\rho_{ij} = ab^{bx_{ij}} + c, \tag{35}$$

where x_{ij} is the distance between locations i and j in kilometers and a, b and c are parameters. The fit for 14 of the 19 low altitude locations can be seen in the Figure 17.

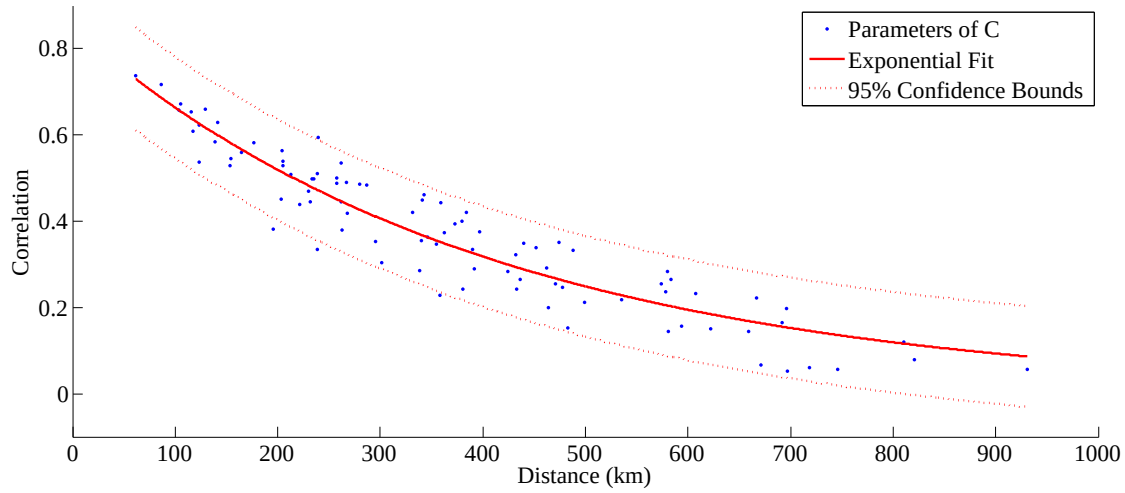

Figure 17: The correlations between different wind generation locations (how hourly wind speed conditions in different generation locations are correlated with each other) as a function of the distance between the locations. The blue markers are the parameters of the correlation matrix $\boldsymbol{C}$ and the red curve is the fit, which used to obtain the correlation between new locations.

After the curve specified by Equation (35) has been fitted to the parameters of the correlation matrix $\boldsymbol{C}$, distances from the new location to all of the existing locations has to be calculated. With the distances know, the new parameters of $\boldsymbol{C}$ can be determined from the fitted curve specified by Equation (35). If conditions in the new locations are similar to the conditions in the measured locations, the average values of the AR parameters of the measured locations can be used for the new locations. Analogously the average values of the monthly diurnal variations in measured locations can be used for the new locations. In addition, the Weibull parameters for the new locations are also required. These parameters can be obtained for the coordinates of the new locations from the the Wind Atlas Database [1].

6 Verification of the Wind Speed Models

This chapter verifies the presented models and is divided into two parts. The first part contains the verification of the transformed VAR model with time-dependent intercept term and the transformed ARC model with diurnal variations for existing locations. The results of 100 Monte Carlo simulation runs from both models are compared against the measurement data from two high altitude locations (Hyytiälä and Puijo) and 14 of the 19 low altitude locations. The locations can be seen on the map of Finland in the Figure 18. The second part presents the verification of the transformed ARC model with diurnal variations for the addition of new locations. 100 Monte Carlo simulation runs from the model for five new locations are compared against the measurement data from these 5 of the 19 low altitude locations, which were not used in the estimation or simulation process in any way.

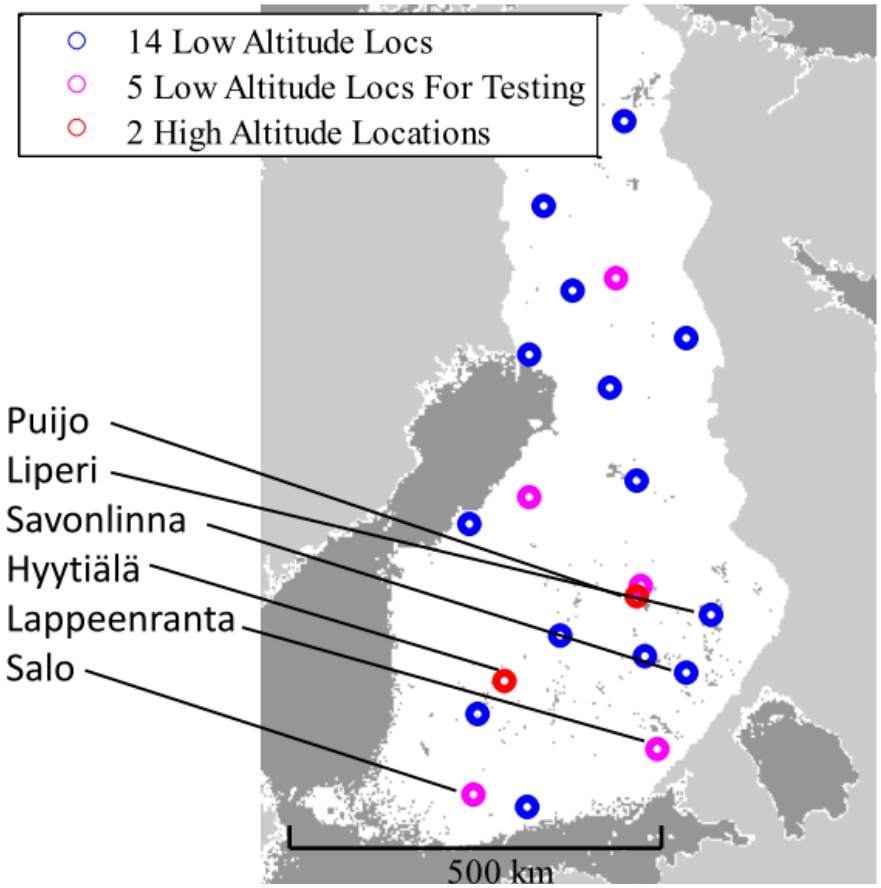

Figure 18: The high and low altitude measurement data locations on the map of Finland. The locations of the 19 low altitude measurements are divided into 14 locations used in the verification of existing locations and estimation of the models and five test locations used in the verification of the addition of new locations. The locations which are used to illustrate the verification results in this chapter are also named in the map.

6.1 Verification on Existing Locations

In this chapter, the Monte Carlo simulation results from 100 simulation runs for the transformed VAR model with time-dependent intercept term and transformed ARC model are compared against the measured data from the two high altitude locations and the 14 low altitude locations. The five locations are left out, because they are considered only in the next section, when the transformed ARC model is verified for new non-measured locations. From the 14 locations, locations Liperi and Savonlinna are used to graphically present the results. Liperi and Savonlinna were chosen because they are relatively close to each other so a reasonably high cross-correlation could be achieved.

This chapter is divided into five parts. First, the marginal distributions are assessed, then autocorrelations, cross-correlations and diurnal variations. Last, few events which require that all of the different aspects presented are modeled correctly are presented and compared against the measurement data.

6.1.1 Marginal Distributions

Figure 19 presents the empirical cumulative density functions (ECDFs) of the measured data from the high altitude locations Hyytiälä and Puijo and Weibull distribution fitted to the measurement data. Figure 19 contains also the marginal distributions of the simulated wind speed time series $\tilde{\boldsymbol{Y}}_i$ for both transformed VAR model with time-dependent intercept term and transformed ARC model. Figure 20 contains the same information for the low altitude locations Liperi and Savonlinna respectively. It can be seen from both of the figures that all four graphs are effectively top of each other, which means that the Weibull distribution fits well for the measurement data in both cases. It is also visible that the simulation results from both models contain correct marginal distributions for all four locations. The same results applied for the other low altitude measurement locations respectively.

Figure 21 illustrates the behaviour of the fitted Weibull distribution and the ECDFs of the high altitude simulation results $\tilde{\boldsymbol{Y}}_i$ for both models at the highest 2.5 % quantiles. Figure 22 presents the same information for the low altitude case. It can be observed that both of the models simulate the data correctly even in the highest quantiles. This is an important feature, as the highest quantiles correspond to the probabilities of the extreme wind speeds, which are interesting phenomena as the storm limit (cut-out speed) can change the generated power of the wind turbine from the maximum value to zero in a very short time period.

The wind speeds in both low altitude locations are absolutely lower than at a real wind turbine heights, which are approximately 100 meters above the surrounding ground level. However, the results obtained from low altitudes apply correspondingly also for higher altitudes as seen from the Figures 19 and 21.

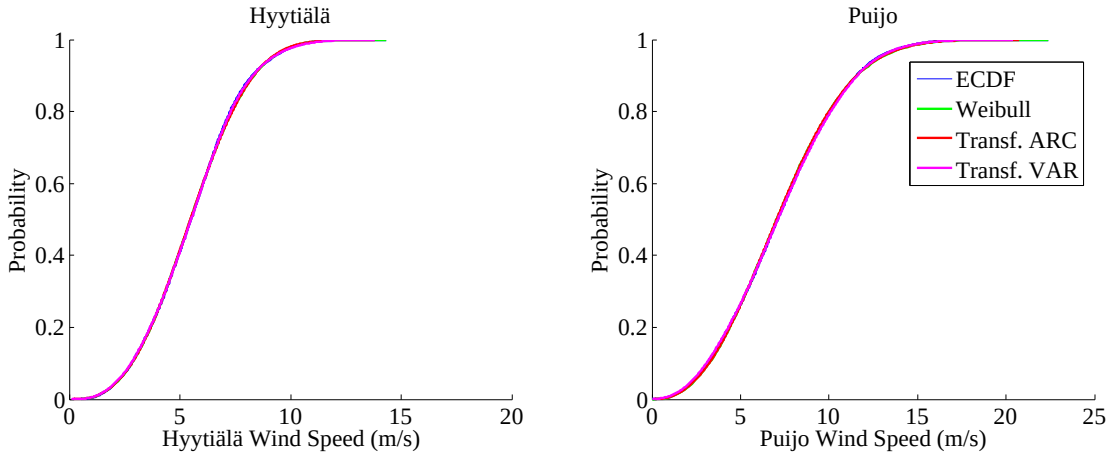

Figure 19: The empirical cumulative distribution function (ECDF) of the measurement data and the average ECDFs of the 100 simulations of the transformed ARC model (Transf. ARC) and the transformed VAR model with a time-dependent intercept term (Transf. VAR) for high altitude Hyytiälä and Puijo.

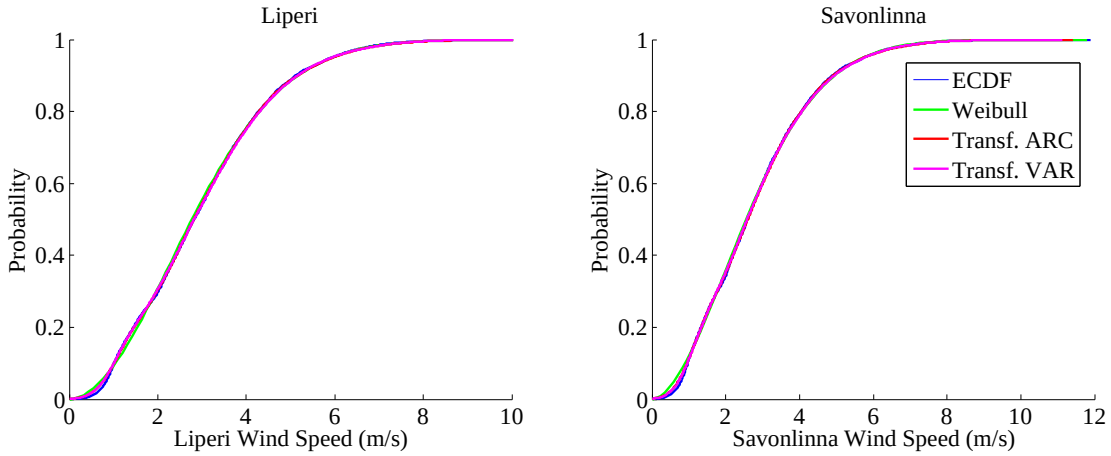

Figure 20: The empirical cumulative distribution function (ECDF) of the measurement data and the average ECDFs of the 100 simulations of the transformed ARC model (Transf. ARC) and the transformed VAR model with a time-dependent intercept term (Transf. VAR) for low altitude locations Liperi and Savonlinna.

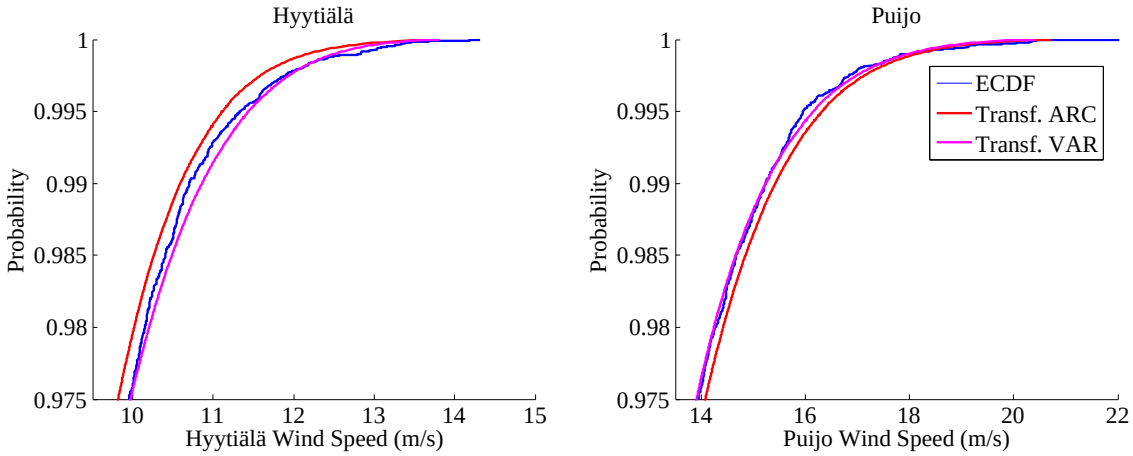

Figure 21: The highest 2.5 % quantiles and the corresponding wind speeds for the empirical cumulative distribution function (ECDF) of the measurement data and the average ECDFs of the 100 simulations of the transformed ARC model (Transf. ARC) and the transformed VAR model with a time-dependent intercept term (Transf. VAR) for high altitude locations Hyytiälä and Puijo.

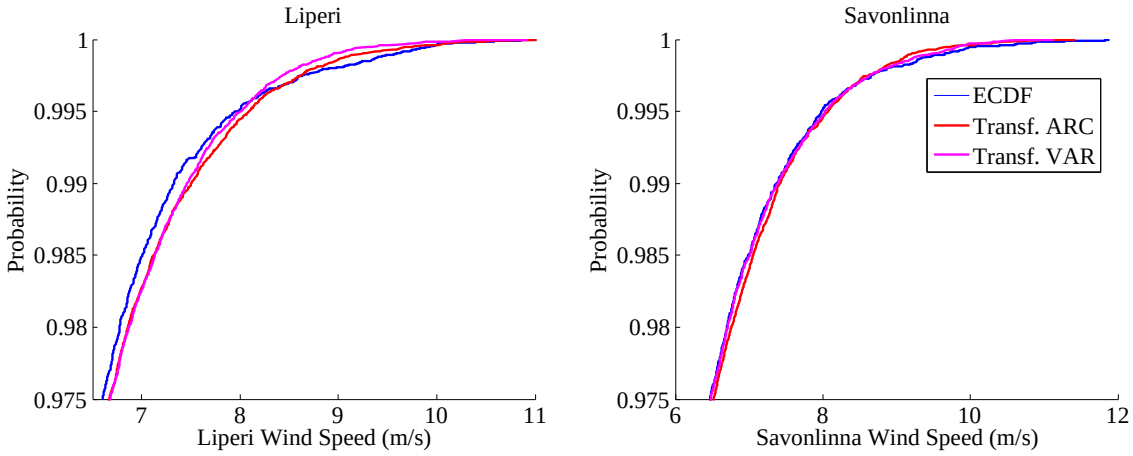

Figure 22: The highest 2.5 % quantiles and the corresponding wind speeds for the empirical cumulative distribution function (ECDF) of the measurement data and the average ECDFs of the 100 simulations of the transformed ARC model (Transf. ARC) and the transformed VAR model with a time-dependent intercept term (Transf. VAR) for low altitude locations Liperi and Savonlinna.

6.1.2 Autocorrelations

Figures 23 and 24 present the ranked autocorrelation functions (RACFs) for the high and low altitude measurement data and both transformed ARC model and transformed VAR model with time-dependent intercept term. It can be observed that both models produce correct RACFs for both the high and the low altitude locations. It can be also observed that the 24-hour bump in the RACFs is not as notable in the high altitude locations as it is in low altitudes as the diurnal variations in wind speeds decrease with higher altitudes. Similar results were obtained for all of the 14 low altitude locations considered.

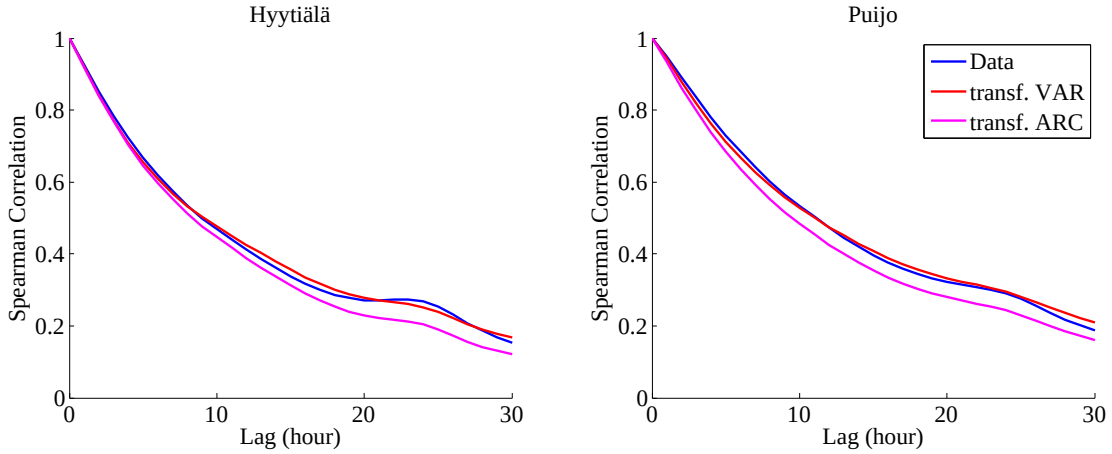

Figure 23: The sample ranked autocorrelation functions (RACFs) of the data and the transformed ARC model and VAR model with time-dependent intercept term for high altitude locations Hyytiälä and Puijo. Autocorrelation is a measure of similarity of time series with itself as a function of a time-lag applied to it. The RACFs are the averages of the 100 Monte Carlo simulation runs.

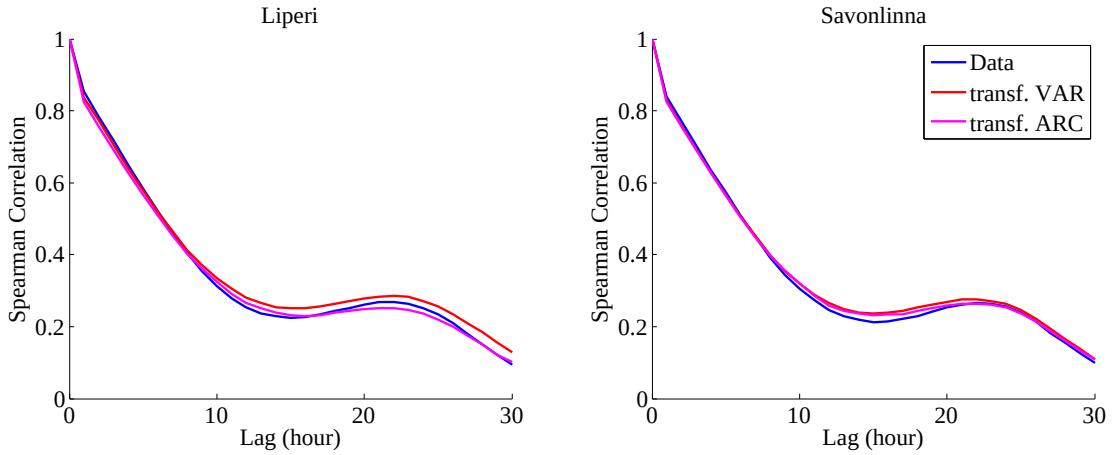

Figure 24: The sample ranked autocorrelation functions (RACFs) of the data and the transformed ARC model and VAR model with time-dependent intercept term for low altitude locations Liperi and Savonlinna. Autocorrelation is a measure of similarity of time series with itself as a function of a time-lag applied to it. The lag in the presented cases in in hours. The RACFs are the averages of the 100 Monte Carlo simulation runs.

6.1.3 Cross-correlations

Figures 25 and 26 illustrate the ranked cross-correlation function (RXCF) for high and low altitude locations. It can be seen that only the transformed VAR model with time-dependent intercept term is able to fully capture the correct shape of the RXCF. The transformed ARC model is not capable of modeling the RXCF correctly, as it assesses the correlation only with lag $h = 0$ between the two locations considered. The transformed VAR model with time-dependent intercept term is able to assess the RXCFs for all h. Graphically the difference between the two models can appear to be significant, but it is shown in Section 6.1.5 that both models are still able to produce good results compared with the data.

Figure 25 also shows that that the transformed ARC model places the peak value always at lag $h = 0$, which, in the case of high altitude locations, is incorrect. Therefore, the transformed ARC is not capable of depicting correctly XCFs where the peak is different from zero. However, this is not a problem with new locations as in that case the peak XCF should be placed at lag $h = 0$ as it is hard to determine any larger patterns for sure (i.e. wind speed direction is mostly from one direction to another) and thus, the only reasonable place for the highest XCF is at lag $h = 0$.

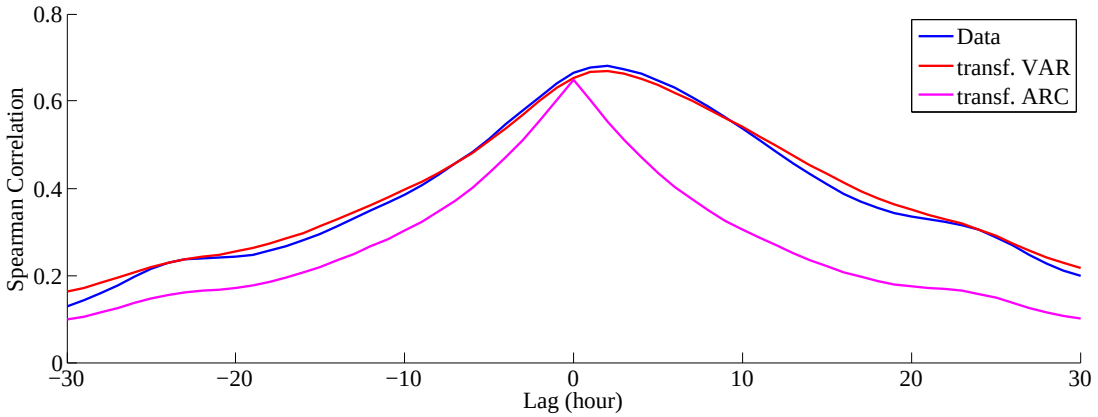

Figure 25: The sample ranked cross-correlation function (RXCF) of Hyytiälä and Puijo for the measurement data and the transformed ARC model and VAR model with time-dependent intercept term. Cross-correlation is a measure of similarity of two time series as a function of a time-lag applied to one of them. The RXCF is the average of the 100 Monte Carlo simulation runs.

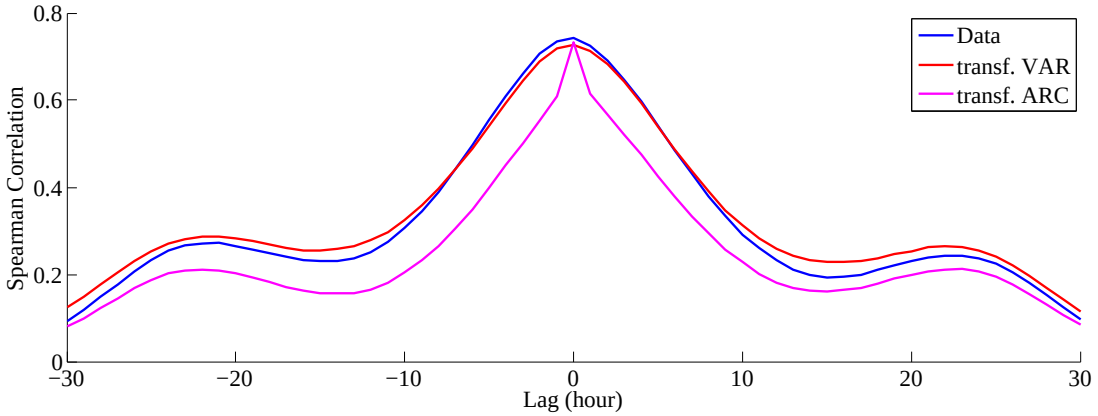

Figure 26: The sample ranked cross-correlation function (RXCF) of Liperi and Savonlinna for the measurement data and the transformed ARC model and VAR model with time-dependent intercept term. Cross-correlation is a measure of similarity of two time series as a function of a time-lag applied to one of them. The RXCF is the average of the 100 Monte Carlo simulation runs.

6.1.4 Diurnal Variations

Figures 27 and 28 present the monthly changing diurnal variation structures of high altitude measurement location Hyytiälä and low altitude location Liperi. It can be observed that both the transformed VAR model with time-dependent intercept term and the transformed ARC model produce correct monthly diurnal structure in the simulated time series $\tilde{\boldsymbol{Y}}_i$ for both high and low locations. The results concerning the diurnal structure were also equally good for all of the 14 low altitude locations.

It can be observed from the Figures 27 and 28 that the monthly diurnal variations differ between high and low altitude locations. The diurnal variations are higher

during summer compared with winter in both altitudes, but the variations are notably larger in low altitude. This is caused by the Sun as it heats the ground at day time during summer which has an effect to the wind conditions. In low altitude, the wind speeds increase during day time because of the heating of the ground and in high altitude the effect is the opposite. It can be also observed that in the high altitude, the higher wind speeds occur during night and in lower altitude during day.

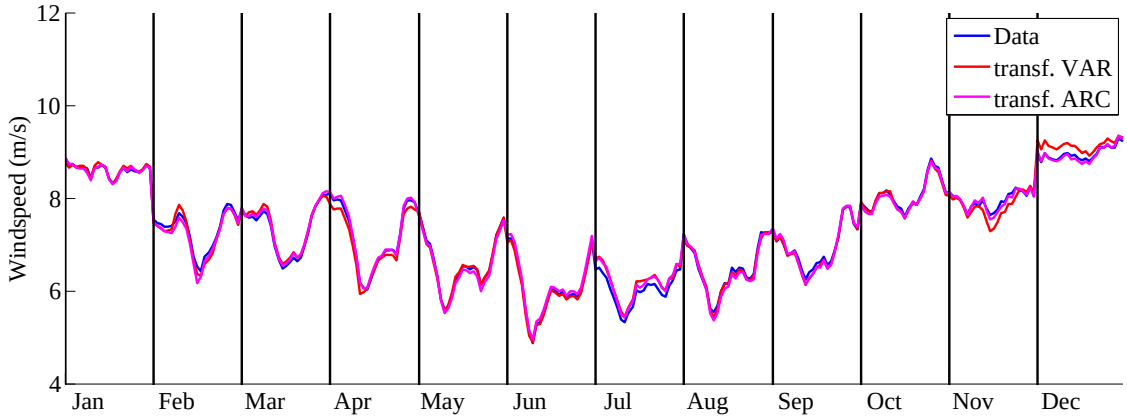

Figure 27: The monthly diurnal variations of the measurement data and the transformed ARC model and the transformed VAR model with time-dependent intercept term for high altitude location Hyytiälä. Each month depicts the average hourly behaviour in the month considered. The presented monthly diurnal variations are the averages of the 100 Monte Carlo simulation runs.

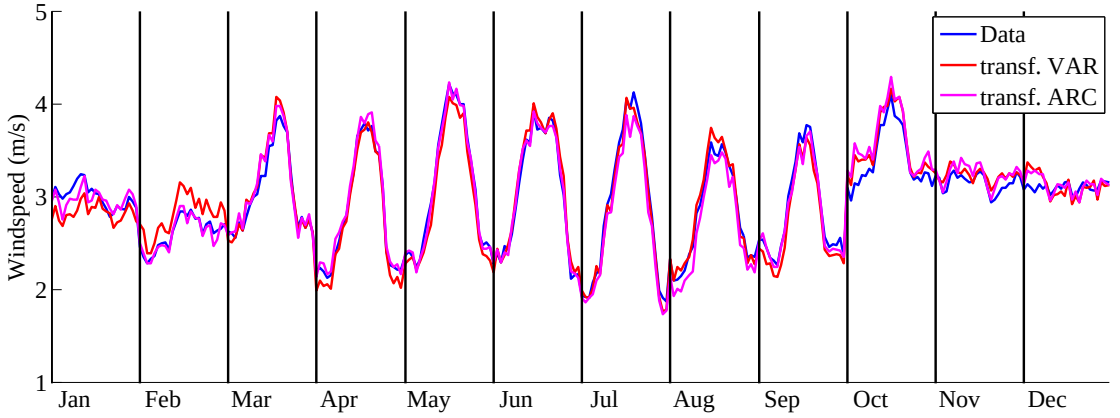

Figure 28: The monthly diurnal variations of the measurement data and the transformed ARC model and the transformed VAR model with time-dependent intercept term for low altitude location Liperi. Each month depicts the average hourly behaviour in the month considered. The presented monthly diurnal variations are the averages of the 100 Monte Carlo simulation runs.

6.1.5 Numerical Verification of Different Wind Speed Events

This section introduces the probabilities of different events for the measurement data from 14 low altitude and two high altitude locations. Table 1 presents the probabilities for several different low and high wind speed events for the low altitude locations and Table 2 presents similar events for the high altitude locations. It can be observed, that both models obtain accurate and very similar results for single location or multiple locations experiencing different wind speed conditions. Thus, both models can be used to estimate the probabilities of different wind speed conditions occurring contemporaneously in several locations.

As the measurement data used in Table 1 is from low altitude locations, the different wind speed events presented are low in comparison with the wind speeds at real wind turbine altitudes. Thus, the scenarios are not relevant if the presented wind speeds are observed. However, Table 1 shows that the models present accurate results compared with the measurement data and if the models work accurately with the low altitude data, they will also work with the high altitude data, as presented with two locations in Table 2, which can be used in the power system risk assessment.

As presented in Section 6.1.3, despite the incapability of the transformed ARC model to correctly model the cross-correlation when $h \neq 0$, the differences in the probabilities presented in Tables 1 and 2 are small. In addition, the differences occur only in the last three rows of Table 1 and last two rows of Table 2 where the wind speed events in consecutive hours are considered. However, if the peak value of the RXCF would exist far from lag $h = 0$ or the shape of the RXCF would be skewed, the transformed ARC model would not give accurate results. Therefore, the transformed VAR model with time-dependent intercept term is a better and more accurate model when analyzing scenarios with existing locations.

Table 1: The probabilities that wind speeds exceed or are less than a given limit in all existing low altitude locations.

The wind speed scenario	Data	Transf. VAR	Transf. ARC
Less than 3 m/s in 5 out of 14 locations	76.60 %	77.19 %	77.17 %
Less than 3 m/s in 10 out of 14 locations	40.87 %	39.79 %	39.67 %
Less than 3 m/s in 14 out of 14 locations	5.57 %	5.62 %	5.69 %
Exceeding 6 m/s in 5 out of 14 locations	6.03 %	5.50 %	5.47 %
Exceeding 6 m/s in 10 out of 14 locations	0.61 %	0.45 %	0.45 %
Exceeding 6 m/s in 14 out of 14 locations	0.06 %	0.01 %	0.01 %
Less than 3 m/s for three consecutive hours in 5 out of 14 locations	24.09 %	24.14 %	23.21 %
Less than 3 m/s for three consecutive hours in 10 out of 14 locations	11.78 %	11.14 %	9.88 %
Less than 3 m/s for three consecutive hours in 14 out of 14 locations	0.97 %	1.01 %	0.85 %

Table 2: The probabilities that wind speeds exceed or are less than a given limit in all existing high altitude locations.

The wind speed scenario	Data	Transf. VAR	Transf. ARC
Less than 6 m/s in 1 out of 2 locations	63.86 %	63.93 %	64.31 %
Less than 6 m/s in 2 out of 2 locations	32.27 %	32.25 %	32.50 %
Exceeding 9 m/s in 1 out of 2 locations	29.86 %	30.50 %	29.57 %
Exceeding 9 m/s in 2 out of 2 locations	5.35 %	4.95 %	4.77 %
Exceeding 12 m/s in 1 out of 2 locations	7.88 %	8.09 %	8.18 %
Exceeding 12 m/s in 2 out of 2 locations	0.19 %	0.19 %	0.12 %
Less than 6 m/s for three consecutive hours in 1 out of 2 locations	19.59 %	19.40 %	19.39 %
Less than 6 m/s for three consecutive hours in 2 out of 2 locations	9.22 %	9.05 %	8.90 %

6.2 Verification on New Locations

In this section, the verification of the transformed ARC model when adding new locations is considered. Again, 100 Monte Carlo simulation runs are made and compared against the measurement data. As presented in Chapter 4, low altitude wind speed measurements were from 19 locations in Finland. 14 of these locations were used in the verifications in previous section and these same 14 locations are also used in this section in the estimation of the transformed ARC model. The five locations, which are not used in any way in the estimation of the model, are used in the verification of the addition of new location to the model.

These five new locations are modeled without any data from the actual measurements in those locations as only the other 14 locations are used. Then, the simulated time series from the five new locations are compared against the actual measurement data. Locations Salo and Lappeenranta were randomly chosen from the five new locations to depict graphically the obtained results.

6.2.1 Marginal Distributions

Figure 29 presents the ECDF of the measurement data and the average ECDF of the 100 simulation runs of the transformed ARC model for Salo, which is one of the new locations. It can be seen from the figure that the marginal distribution of the simulated wind speed time series $\tilde{\boldsymbol{Y}}_i$ produced by the transformed ARC model is close to the actual measurement data, even though the Wind Atlas Database of The Finnish Meteorological Institute (FMI) could not be used with the low altitude measurements as the Wind Atlas provides Weibull parameters only for higher altitudes. Instead, the average Weibull parameters from the 14 location used in the estimation was used for the new locations. Corresponding good results were obtained also for the other four new locations. In case of measurements from higher altitudes, the Wind Atlas Database can be used to obtain correct Weibull parameters for the new locations as presented in Section 5.3.3.

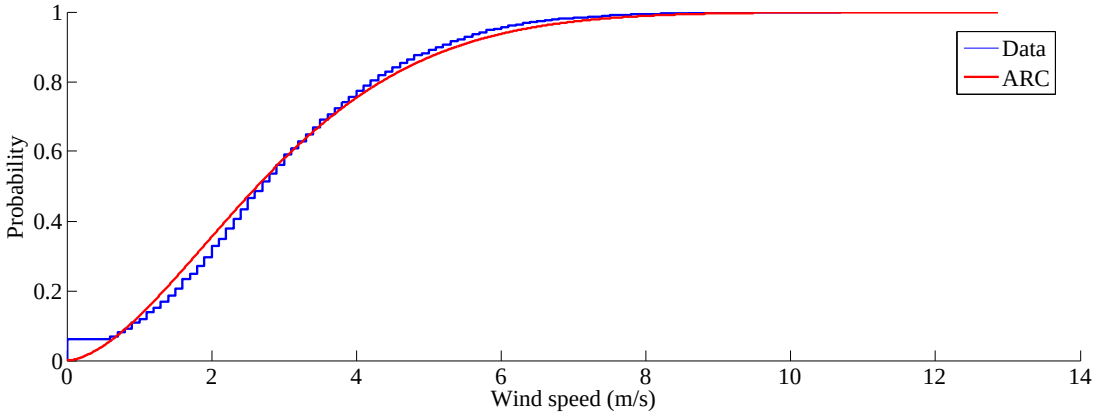

Figure 29: The Empirical cumulative distribution function (ECDF) of the measurement data and the average ECDF of the 100 simulations of the transformed ARC model (ARC) for Salo.

6.2.2 Autocorrelations

Figure 30 presents the ranked autocorrelation functions (RACFs) for the measurement data and the transformed ARC model for new locations Salo and Lappeenranta. As Figure 30 presents, the RACFs obtained for both locations are close to the measurement data. Therefore, using average AR parameters of the existing 14 locations for new locations yield accurate results and can be considered as a solid approach in this case. However, this approach requires that the wind speed conditions in the new locations do not differ notably from the locations used in the estimation of the AR parameters. Similar accurate results were acquired also for the other three new locations.

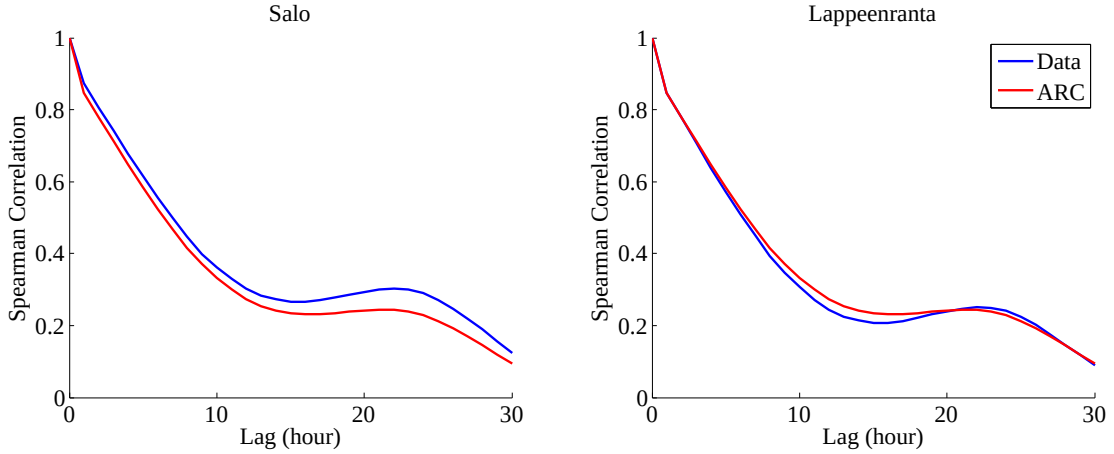

Figure 30: The sample ranked autocorrelation functions (RACFs) of the data and the transformed ARC model for Salo and Lappeenranta. Autocorrelation is a measure of similarity of time series with itself as a function of a time-lag applied to it. The RACF of the transformed ARC model is the average of the 100 Monte Carlo simulation runs.

6.2.3 Cross-correlations

Figure 31 illustrates the ranked cross-correlation function (RXCF) for Liperi and Savonlinna. It can be observed that the transformed ARC model can estimate the RXCF relatively approximately also for the new locations, when the average AR parameters of the existing 14 locations are used.

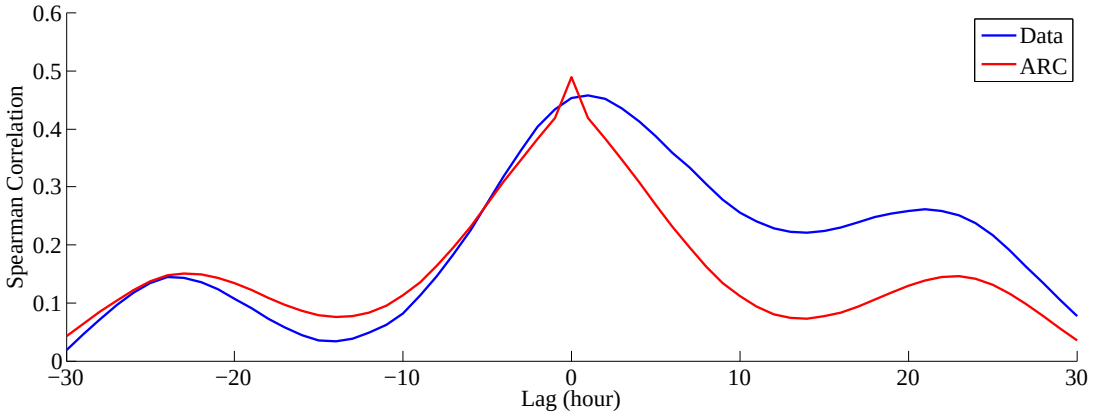

Figure 31: The sample ranked cross-correlation function (RXCF) of Salo and Lappeenranta for the measurement data and the transformed ARC model. Cross-correlation is a measure of similarity of two time series as a function of a time-lag applied to one of them. The RXCF is the average of the 100 Monte Carlo simulation runs.

6.2.4 Diurnal Variations

Figure 32 presents the monthly changing diurnal variation structure of Salo, which is one of the new locations. As shown in Figure 32, the transformed ARC model produces correct monthly diurnal structure for Salo. Corresponding accurate fits were also obtained for the other four new locations. Therefore, the usage of the average monthly diurnal variation structure of the existing 14 locations for the new locations result in accurate results and can be considered as a reliable approach in these conditions. However, this approach has the same requirements as the usage of the average AR parameters and it does not allow a huge variation in the conditions in the existing locations used in the estimation process and in the new locations added to the model.

In addition, it can be seen from Figure 32 that the diurnal variations are higher during summer compared with winter. This is caused by the Sun as it heats the ground at day time during summer which increases the wind speed conditions in low altitudes. During night when the ground cools, also the wind speeds are reduced.

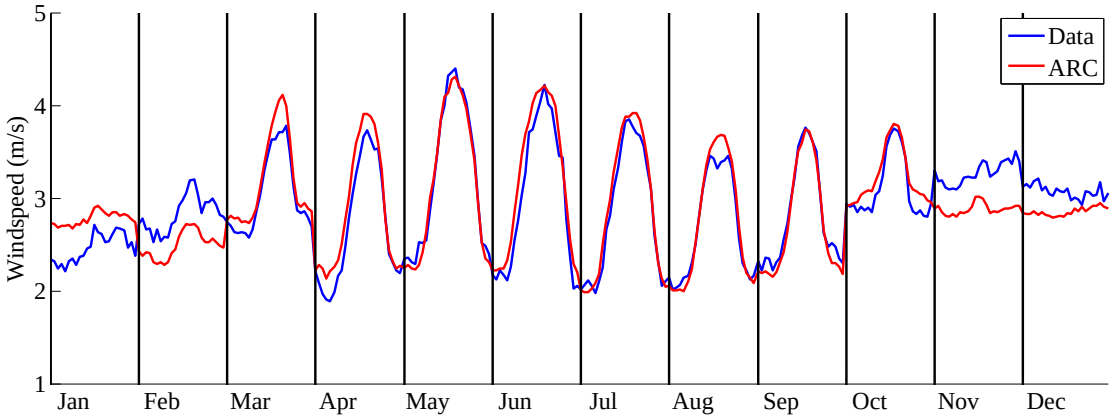

Figure 32: The monthly diurnal variations of the measurement data and the transformed ARC model for Salo. Each month depicts the average hourly behaviour in the month considered. The presented monthly diurnal variations are the averages of the 100 Monte Carlo simulation runs.

6.2.5 Numerical Verification of Different Wind Speed Events

In this section, the probabilities for different wind speed events for the measurement data from the five new locations are compared with the simulation results when adding new locations to the model. Table 3 illustrates the probabilities for several different low and high wind speed events for the new locations.

As in Section 6.1.5 with Table 1, the measurement data used in Table 3 is from low altitude locations and therefore the different wind speed events presented are low in comparison with the wind speeds at real wind turbine altitudes. Thus, the scenarios are not relevant if only the presented wind speeds are observed. However, Table 3 shows that the models present accurate results with new non-measured locations compared with the measurement data and if the models work accurately with the low altitude data, they will also work with the high altitude data, which can be used in the power system risk assessment. Therefore, the transformed ARC model is suitable for the analysis of low or high wind speeds occurring contemporaneously in multiple locations in scenarios with new non-measured locations. In the next chapter, different cases where the transformed ARC model is used with non-measured locations are presented.

Table 3: The probabilities that wind speeds exceed or are less than a given limit in all new locations.

The wind speed scenario	**Data**	**Transf. ARC**
Less than 3 m/s in 1 out of 5 locations	90.94 %	90.81 %
Less than 3 m/s in 3 out of 5 locations	62.28 %	60.24 %
Less than 3 m/s in 5 out of 5 locations	21.31 %	20.02 %
Exceeding 6 m/s in 1 out of 5 locations	18.28 %	21.37 %
Exceeding 6 m/s in 3 out of 5 locations	2.32 %	2.38 %
Exceeding 6 m/s in 5 out of 5 locations	0.07 %	0.12 %
Less than 3 m/s for three consecutive hours in 1 out of 5 locations	29.37 %	28.85 %
Less than 3 m/s for three consecutive hours in 3 out of 5 locations	18.43 %	16.67 %
Less than 3 m/s for three consecutive hours in 5 out of 5 locations	4.94 %	4.09 %

7 Results on New Wind Power Scenarios

In this chapter, results obtained from four example cases, introduced in Table 4, are presented, assessed and compared against the data from Finland. The results are obtained with the transformed ARC model estimated with 19 low altitude locations from Finland. 100 Monte Carlo simulation runs have been made for each case to obtain the results. The four cases presented can be seen on the map of Finland in Figure 33. In every case 10 wind farms are placed to new non-measured locations. Each location consists of 10 Vestas V205 - 3.3 MW IEC IA wind turbines (turbine data is available in [27]) with nominal power of 3.3 MW and height of 100 m resulting in total generation capacity of 330 MW in each case. The 10 turbines in one wind farm are considered as one big generation unit in specific coordinates.

Table 4: The specifications of the four cases assessed in this chapter.

	Short distances	Long distances
High altitude Weibull parameters	Case 1	Case 2
Low altitude Weibull parameters	Case 3	Case 4

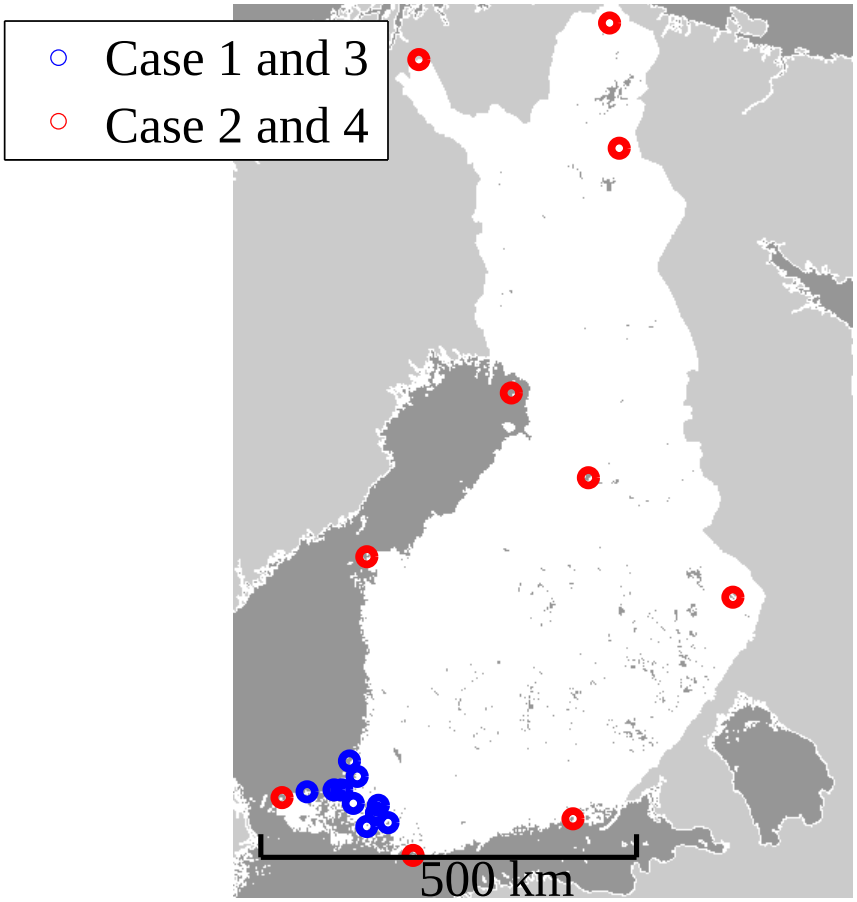

Figure 33: The map of Finland depicting the locations of the wind farms in the four example cases.

7.1 Conversion from Wind Speed to Power

The conversion from wind speeds to wind power is not straightforward if it is analyzed precisely. However, in the four example cases presented in Chapter 7, the conversion is linear between the cut-in and the nominal speed of the wind turbines. The conversion can be seen from Figure 34. The presented power curve comes from the location 10 in Case 1. The distribution of the output power, which can be seen next to the y-axis, shows peaks at zero power and nominal power because only one wind turbine is considered. The peaks will vanish if the aggregate power of multiple turbines with high Weibull parameters in each turbine location and long distances (or low correlation) between the locations are considered.

The presented power curve cannot be verified as no single turbine power data is available. However, the obtained curve is created with the parameters of Vestas V205-3.3 MW IEC IA wind turbine [27], which is the turbine type that is also used in the presented cases.

In these example cases, a linear power curve is used, however, it is straightforward to implement any kind of turbine specific power curve to the model. Also, the hysteresis effect of the power curve i.e. the changed cut-in speed after the cut-out in the maximum speed is discussed in Section 8.3.

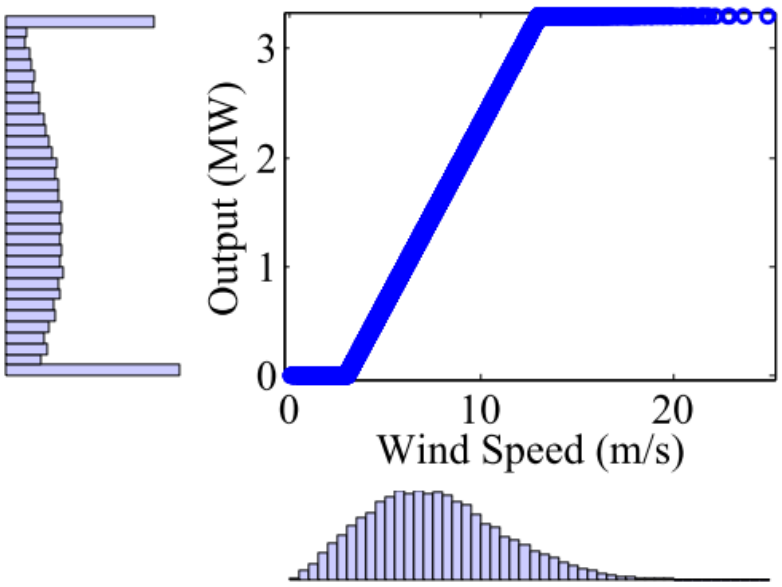

Figure 34: The power curve from the location 10 in Case 1 (Weibull parameters $A = 8.6$ and $B = 2.3$) depicting the transformation from the wind speeds to wind power. The distribution of the wind speeds can be seen below the x-axis and the distribution of the output power next to the y-axis.

7.2 Simulating Different Cases with New Locations

Next, the four cases presented in Table 4 are introduced. For Case 1, 10 wind farms are placed close to each other in the southwestern archipelago of Finland. The high altitude (100 meters above the sea level) Weibull parameters for each location are obtained from the Wind Atlas Database [1]. In Case 2, 10 wind farms are placed far from each other to different location in Finland from the southwestern archipelago of Finland to the most northern parts of the country as seen in the Figure 33. In Case 2, the high altitude (100 meters above the sea level) Weibull parameters are also obtained from the Wind Atlas Database. To obtain results which can be easily compared with each other, the locations have been chosen so that the Weibull parameters are very close to each other in both cases despite the fact that the parameters are not fixed but real values. In Case 3, the locations of the wind farms are same as in the Case 1, but now the Weibull parameters have been artificially reduced to values $A = 6$ and $B = 2$ to illustrate the effect of smaller Weibull parameters or lower altitudes i.e. smaller wind speeds in general for aggregated wind power generation. In Case 4, the locations are same as in the Case 2, but the Weibull parameters are reduced to $A = 6$ and $B = 2$ as in Case 3. In all cases, the aggregate generation capacity for all of 10 wind farms is 330 MW.

7.2.1 Histograms

Figure 35 presents the histogram of the aggregate hourly wind power generation of Case 1, presented in Table 4 and Figure 33, in time period equal to the length of the estimation data from 19 low altitude locations (T = 23735). It can be seen from the Figure 35 that the histogram is flat with a small peak at the maximum generation. There is no peak at zero generation as the Weibull parameters in all of 10 wind farm locations are high i.e. the locations have very high general wind speeds.

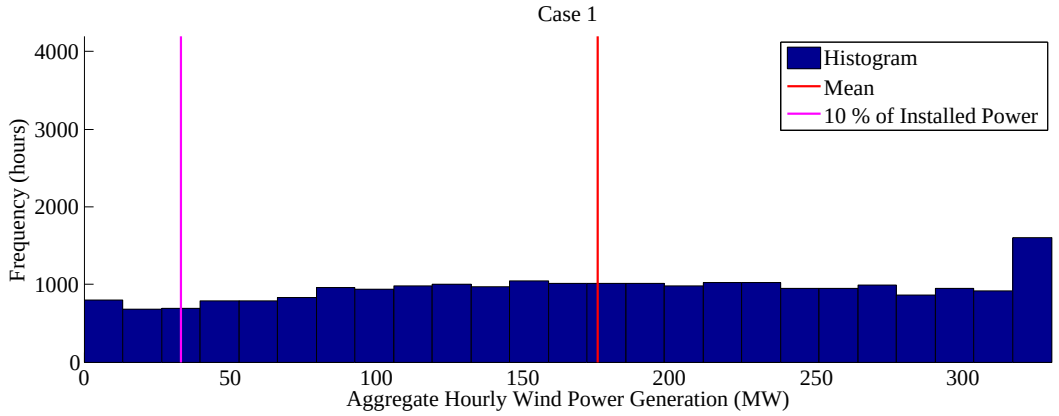

Figure 35: The histogram of the aggregate hourly wind power generation in Case 1.

Figure 36 presents the histogram of the aggregate hourly wind power generation in Case 2, presented in Table 4 and Figure 33. It can be observed that the distribution

of the histogram has the most of its values near the mean value. This is due to the long distances, and thus small correlation, between wind farms. Also, as the wind speed conditions are similarly high as in Case 1, there is no peak in the zero generation. Case 2 is the most favorable situation from the viewpoint of the power system operator as the generation is rare near zero or nominal value. It can be also observed that the mean values in Cases 1 and 2 are almost the same because the Weibull parameters of the wind farm locations are very close to each other in both cases. The mean value in Case 1 is slightly higher (175 MW) than in Case 2 (169 MW), however as the difference is small, the cases can be easily compared with each other. As the mean values are close to each other, so are the yearly aggregate generations. We obtain the yearly generation of 1536 GWh for Case 1 and 1478 GWh for Case 2.

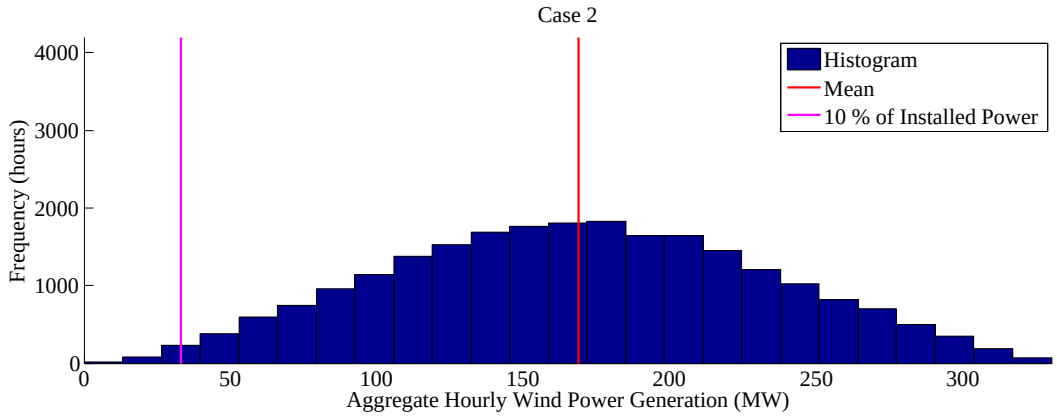

Figure 36: The histogram of the aggregate hourly wind power generation in Case 2.

Figure 37 presents the histogram of the aggregate hourly wind power generation in Case 3, presented in Table 4 and Figure 33, where the location of the wind farms is identical to Case 1. However, Case 3 illustrated the situation where the general wind speed conditions in wind farm locations are lower i.e. the Weibull parameters are smaller, scale parameter $A = 6$ and shape parameter $B = 2$. Now, the peak at zero generation is clearly visible due to the lower wind speeds in general.

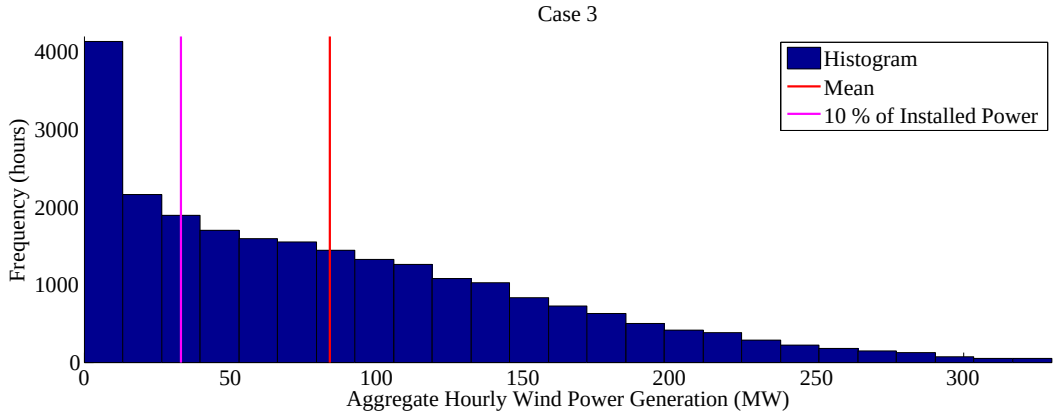

Figure 37: The histogram of the aggregate hourly wind power generation in Case 3.

Figure 38 presents the histogram of the aggregate hourly wind power generation in Case 4, presented in Table 4 and Figure 33. The locations are identical to Case 2, but the Weibull parameters are lowered to $A = 6$ and $B = 2$ as in Case 3. In Cases 3 and 4, where the Weibull parameters are lower, the mean values are also very close to each other. The mean values are 84 MW for Case 3 and 81 MW for Case 4. The yearly aggregate generation is 734 GWh for Case 3 and 707 GWh for Case 4.

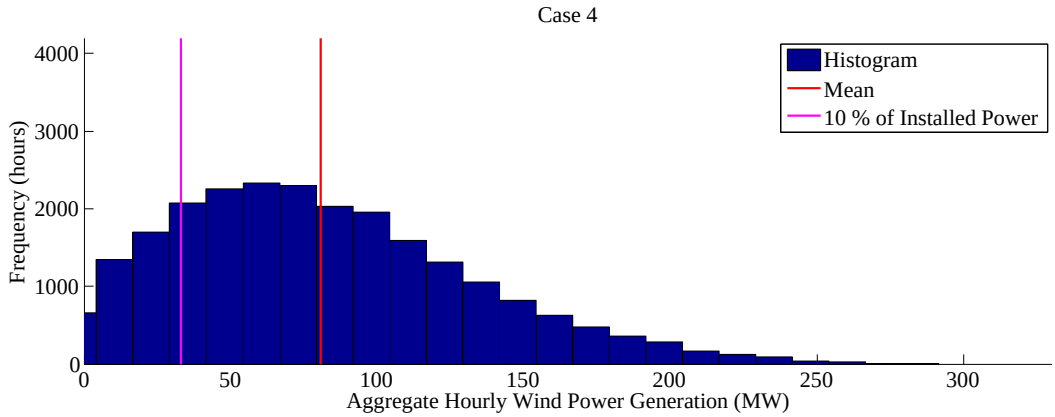

Figure 38: The histogram of the aggregate hourly wind power generation in Case 4.

For operation planning, it is useful to estimate the probabilities of situations where the generation is lower or higher than a certain limit. Let us consider generation less than 10 % of the wind generation capacity as an example. If the percentage of hours when generation is less than 10 % of the aggregated wind generation capacity is observed, the following notions can be made. In Case 1 with short distances between the wind farms and high Weibull parameters, generation is less than 10 % of the capacity in 7.6 % of hours during the simulation period. In Case 2, where the distances between the wind farms are long and the Weibull parameters are high, the generation is less than 10 % of the capacity only in 0.8 % of the hours. In Case 3, with short distances and lower Weibull parameters, the percentage is 30,6 %. In Case 4 with long distances and lower Weibull parameters, the percentage is 18,6 %.

Therefore, as visible in Figures 35, 36, 37 and 38, the most favorable option for the power system operator is Case 2, where the correlations between wind farms are small and the Weibull parameters are high. In Case 2, the power balance in the power system is easiest to maintain as the events with very high or low generation are rare.

Figure 39 presents the histogram of the aggregate hourly wind power generation in Finland between 2008 and 2012. It can be observed that the histogram resembles a combination of Cases 3 and 4. However, the geographical spread and the aggregate generation capacity have both grown during the considered time period, which can have an effect on the shape of the histogram.

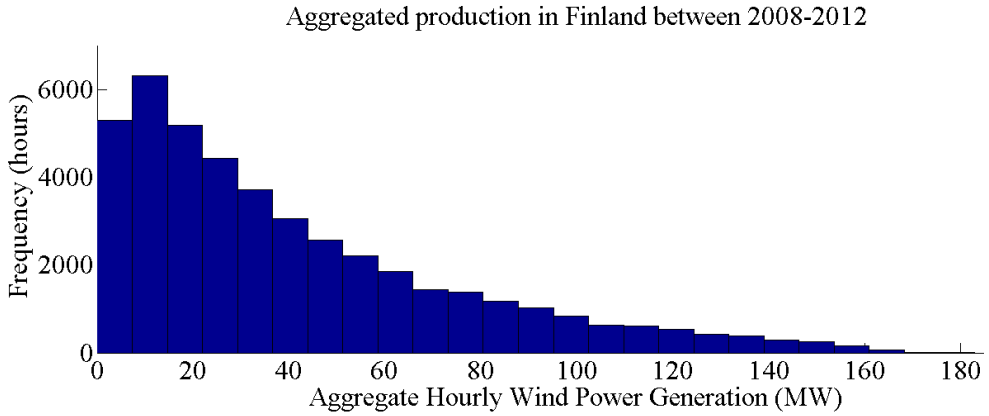

Figure 39: The histogram of the aggregate hourly wind power generation in Finland between 2008 and 2012.

7.2.2 Autocorrelations and Time Series

Figure 40 presents the autocorrelations of the aggregate power generation in the four cases. It can be observed that the autocorrelations are identical in Cases 1 and 3 (short distances between wind farm locations in both cases), and in Cases 2 and 4 (long distances between wind farm locations in both cases) as only the Weibull parameters change which does not affect the autocorrelations. In cases 2 and 4, the 24-hour day structure is also more notable than in Cases 1 and 3. It can be useful to observe the autocorrelations of the aggregate generation as they are able to depict how likely it is that the generation is close to current generation after a certain time interval.

Figure 41 presents the autocorrelation of the aggregate power generation in Finland between 2008 and 2012. It can be seen that the autocorrelations in Finland differ notably from the autocorrelations of the four example cases. There is no notable day structure visible in the Figure 41 and the autocorrelations are much higher than in Figure 40. This difference is due to the fact that the model used in the simulation of the four example cases was estimated with low altitude measurements. In lower altitudes, the autocorrelations are lower and the effect of the day structure is more

remarkable than in higher altitudes. However, these problems with autocorrelation can be avoided if the estimation of the model is done with high altitude measurements. In the case of the four example cases assessed in this chapter, the estimation of the model was done with the low altitude measurements as high altitude measurements were available only from two locations, which is not enough for the accurate estimation of the model.

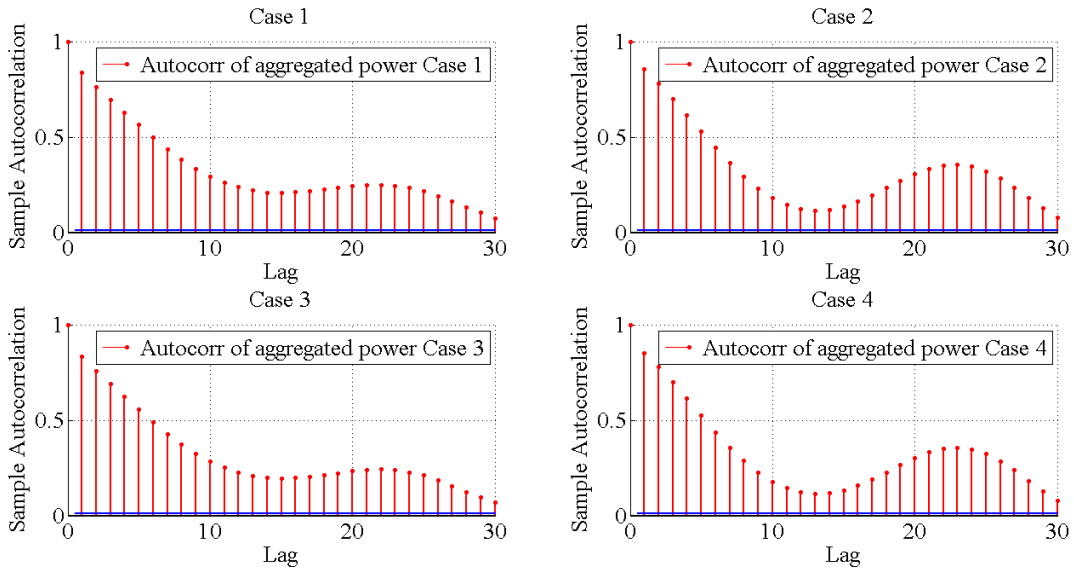

Figure 40: The autocorrelations of aggregate power generation in the four example cases. Autocorrelation is a measure of similarity of time series with itself as a function of a time-lag applied to it. The lag in the presented cases is in hours.

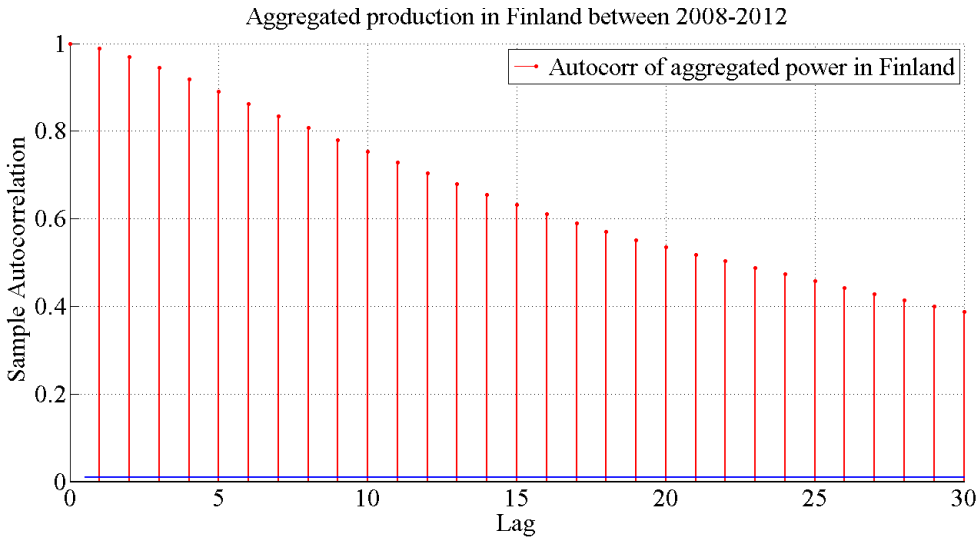

Figure 41: The autocorrelations of aggregate power generation in Finland between 2008 and 2012. Autocorrelation is a measure of similarity of time series with itself as a function of a time-lag applied to it. The lag in the figure is in hours.

Figure 42 presents the time series of the aggregate power generation in the four

cases. It can be observed that in Case 1 the time series gets both zero and maximum values constantly. In Case 2 the time series gets both zero and maximum values very rarely. In Case 3 the time series gets constantly zero values but very rarely maximum values. In Case 4 the time series gets less constantly zero values compared with Case 3 but more often than in Case 2 and never maximum values. All these observations correspond well with the histograms considered earlier in this chapter.

If the time series of the four cases are compared with the time series of the aggregated generation in Finland presented in the Figure 2, it can be observed that the time series of Case 4 is the closest match. Both time series get zero values frequently but never maximum values. These observations are similar to those obtained from the comparison of the histograms of the four cases and the corresponding data from Finland.

The visual observation of the time series can be useful as additional verification that the transformed ARC model produces correct time series structures. In addition, the time series can present more information if they are observed with different time intervals.

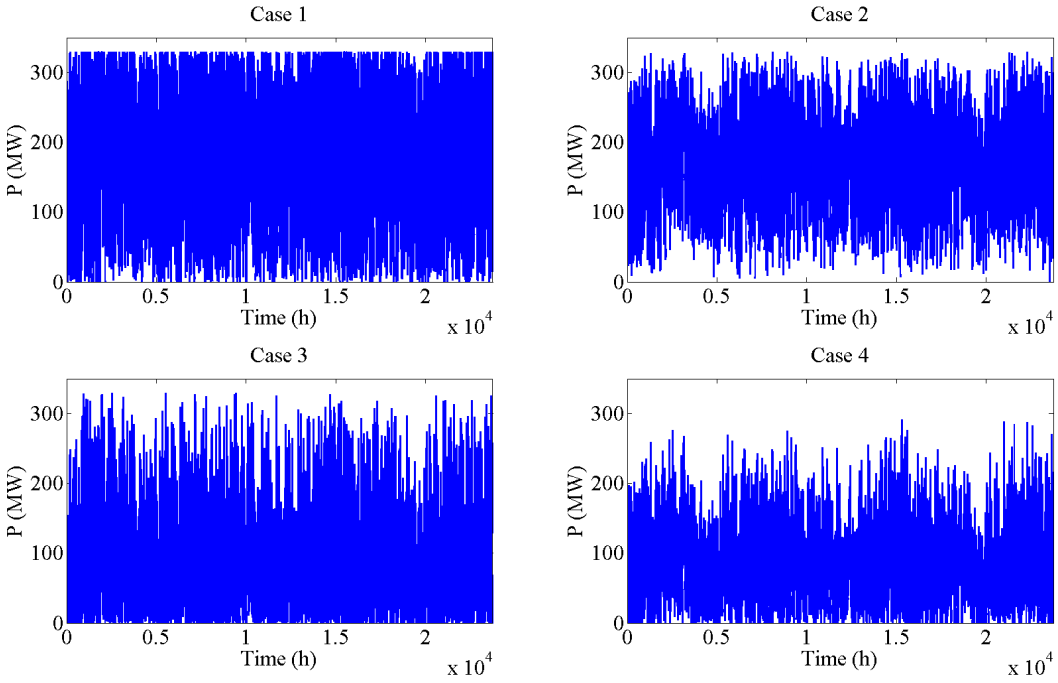

Figure 42: The time series of the aggregate power generation in the four cases.

7.2.3 Temporal Dependency Structures

Figure 43 presents the temporal dependency structures between the aggregate generation at time t and $t-1$. It can be observed that the clouds of points in Figure 43 depicting Cases 2 and 4, where the correlations between wind farms are low, are

narrower compared with Cases 1 and 3. This is due to the lower correlation between wind farms. When the correlations between locations are lower, the changes in aggregate generation are also smaller as situations where the generation in multiple wind farms would increase or decrease simultaneously are rarer. This kind of assessment of the temporal dependency structures can be useful, as it is able to illustrate the possibility of major changes in aggregate generation during a certain time interval.

Figure 44 presents the temporal dependency structure of the aggregate wind power generation in Finland between 2008 and 2012. It can be observed that the dots form a relatively thin cloud of points. Compared with the four example cases, the closest resemblance is with Case 4 as was also with the histograms and time series. The thinner shape of the cloud of points in Figure 44 is the result of the higher autocorrelations in the data from Finland compared with the four example cases. Therefore, if the simulation model is estimated with high altitude data, also the temporal dependency structures in Figures 43 and 44 would resemble more closely each other.

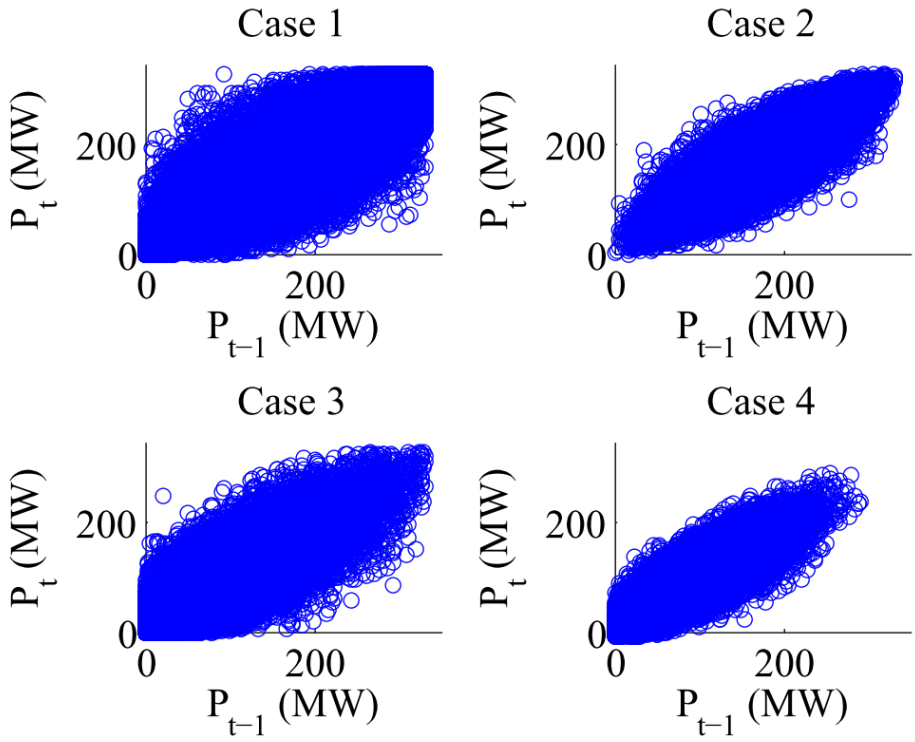

Figure 43: The temporal dependency structures of aggregate power generation between t (y-axis) and $t-1$ (x-axis) i.e. one hour in the four example cases with new locations.

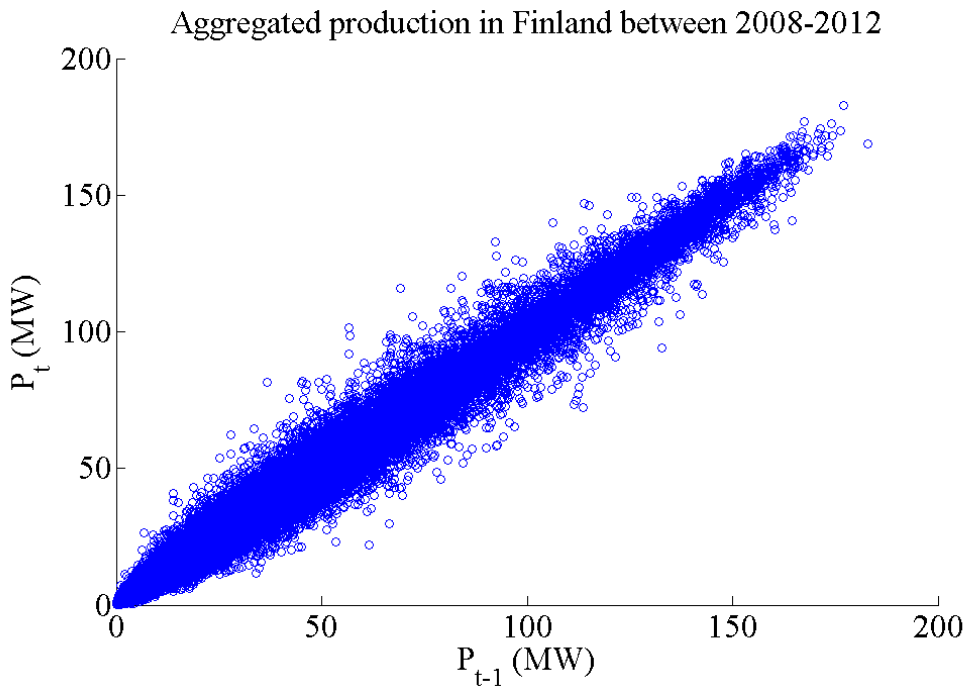

Figure 44: The temporal dependency structure of aggregate power generation between t (y-axis) and $t-1$ (x-axis) i.e. one hour in Finland between 2008 and 2012.

8 Discussion

This chapter presents a few matters that require more discussion. First, Archimedean copulas and t-copula are shortly discussed as these copulas were not considered in Chapter 3. Second, the applicability of the models presented in this thesis is discussed. Last, the scope of future research related to this thesis is discussed.

8.1 Archimedean Copulas and t-copula

In this thesis, it is assumed that the Gaussian copula specifies the dependency structures between two variables (both spatial and temporal dependencies). However, there are numerous other copulas, which could be also considered as mentioned in Section 3.6. In this section, t-copula and the Archimedean family of copulas are shortly discussed.

As Gaussian copula is based on a multivariate normal distribution, the t-copula is based on a multivariate t-distribution [9]. As Gaussian copula has only one parameter, covariance matrix $\boldsymbol{\Sigma}$ equal to the correlation matrix $\boldsymbol{C}$, t-copula has two, correlation matrix $\boldsymbol{C}_t$ and the degree of freedom v of the t-distribution [21]. t-copula will always give a better fit compared with Gaussian copula as it has one extra parameter v and as Gaussian copula is a special case of t-copula when $v \to \infty$.

In this thesis, t-copula is not assessed as copulas with more than one parameter were not considered. If t-copula would be implemented, it would increase the complexity of the models but bring only slight improvements compared with Gaussian copula.

As Gaussian and t-copula are derived from probability distributions, the family of Archimedean copulas can be stated directly. The Archimedean copulas can be written as

$$C(\boldsymbol{u}) = \phi(\phi^{-1}(u_1) + \phi^{-1}(u_2) + ... + \phi^{-1}(u_k)), \tag{36}$$

where ϕ is called the generator. The generator is different for different Archimedean copulas. Commonly used Archimedean copulas are Clayton, Frank and Gumbel copulas and now Frank copula is considered as an example.

Figures 45 and 46 illustrate an examples of Frank copula from Archimedean family. Figure 45 presents the Frank copula with normal margins and Figure 46 with Weibull margins. The Frank copula parameter used in both cases corresponds to the Pearson's correlations used in Figure 6. As visible in the Figures 45 and 46, the Frank copula yields very different results compared with the Gaussian copula presented in the Figure 7. Therefore, a copula, which fits best to the empirical data, should be used to depict the dependency structure of the data. In case of this thesis, Gaussian copula fit best if only copulas with one parameter were considered and therefore it was justified to use it.

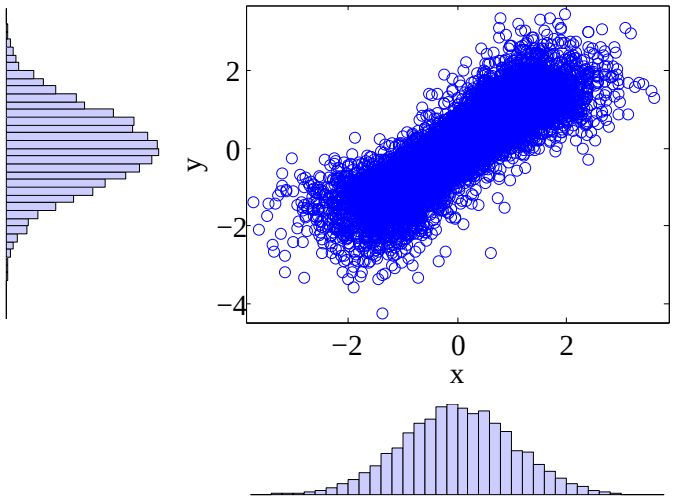

Figure 45: An example of a Frank copula with Normal distributions as margins.

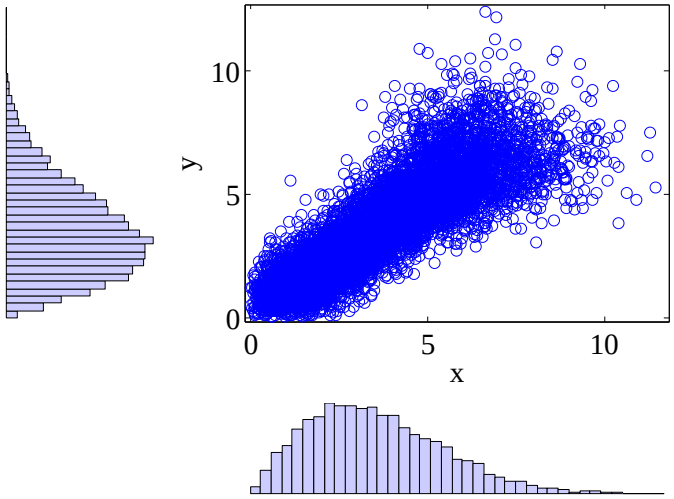

Figure 46: An example of a Frank copula with Weibull distributions as margins.

8.2 Applicability of the Models

The models presented in this thesis can be used for the analysis of wind power generation in several locations. The transformed VAR model with time-dependent intercept term can be used to assess scenarios with existing locations with high accuracy and the transformed ARC model to assess scenarios with new non-measured locations. For non-measured locations, wind speed conditions for a specific location can be determined by using the Weibull parameters obtained from the Wind Atlas Database. The analysis of situations of this kind can be useful when assessing the effect of wind farms still in the planning stage on the power system.

The presented statistical models can be used also in the analysis of the variation in

wind power generation when more generation is implemented in the system. The increased generation can be analyzed both in new locations or the as increased capacity in the current generation locations or combination of these two.

As both of the models use wind speeds instead of wind power, different wind turbine types and power curves specific to each turbine for every location can be easily implemented into the analysis. This enables straightforward simulation of scenarios with different wind turbine types. Through this link from wind speeds to wind power, the presented models can be used as a tool for power system planning.

It is possible to calculate probabilities for extreme situations (near zero wind speeds or wind speeds over the cut-out speed of the turbines) in multiple locations simultaneously. Probabilities can be calculated also for aggregate generation exceeding or being less than a certain percentage of the installed capacity. Also, the probabilities for the change of wind power generation within a certain time period exceeding a certain value can be calculated. Analyses of this kind enable the system operator to prepare against different generation scenarios.

The analyses can be used in the long term planning of the power systems with high penetration of wind generation. The extreme situations can be analysed and the results of the analyses can be used in the determination of the sufficient amount of balancing power for maintaining the power balance in the system.

8.3 Future Work

This section considers the planned future work concerning the models presented in this thesis. Next, the most relevant focus areas of the future research are considered.

The transformed VAR model with time-dependent intercept term may be possible to utilize for the addition of new locations through Yule-Walker equations. This needs to be considered in the future as the transformed VAR model with time-dependent intercept term is a more accurate model than the transformed ARC model.

New high altitude data has been acquired from multiple locations in Finland and it will be used in the analysis of different and more detailed wind power generation scenarios. Also, the power curve used in the transformation of wind speed to wind power will be extended with the hysteresis effect (different cut-in speed after the shutdown due to wind speeds higher than the storm limit) and a nonlinear curve.

Research concerning wind speed and power simulations extended from the models introduced in this thesis is planned. The planned research includes the modeling of the whole wind power generation structure in Finland and in future also in the Nordic countries, short-term wind speed and power forecasts for the use of the system operator, and more accurate models with t-distributed error terms and implementation of the generalized autoregressive conditional heteroskedasticity (GARCH) model, which implements changing variance to the models, for power system planning. In

addition to the wind speed and power, one goal is also to implement other types of renewable generation i.e. solar to the models.

9 Conclusions

The main objective of this thesis was to develop two simulation models for the assessment of large scale wind power generation and evaluate the feasibility of the models. The transformed VAR model with time-dependent intercept term and the transformed ARC model were developed for the wind speed simulations and the feasibility of the models was assessed. Therefore, the main objective of this thesis was successful.

The presented models are used in Monte Carlo simulations to determine the risks of very high or low wind speeds occurring in several locations simultaneously. The models combined copula modeling to autoregressive models. The benefit of copula modeling was that the marginal distributions of each location could be separated from the dependency structures. This allowed the complete analysis of the dependency structures, which otherwise would have been difficult.

The models were verified for existing locations against measurement data from 19 low altitude and two high altitude locations in Finland. The results showed that the transformed VAR model with time-dependent intercept term was slightly more accurate than the transformed ARC model. It was able to capture the full spatial and temporal dependency structures as the transformed ARC model was not capable of depicting the structure of the cross-correlation functions (XCFs) correctly with non-zero lags. However, both models gave good results compared with the measurement data, and therefore allowed the assessment of the extreme events in wind power generation occurring simultaneously in several locations.

The transformed ARC model was the only model considered when adding new non-measured locations as it provides an accurate and straightforward method for the implementation of new locations to the simulations. The transformed VAR model with time-dependent intercept term was not considered as a feasible approach when adding new locations to the simulations as it had major problems with the implementation of the new non-measured locations. The transformed ARC model was verified for new locations against five low altitude locations in Finland, which had not been used in any way in the estimation of the model. The results were accurate compared with the measurement data from these locations, and thus the model can be used in scenarios where new non-measured locations are added to the simulations.

The wind speeds, and therefore the marginal distributions were considered Weibull distributed. This is an advantage in power system planning as the Wind Speed Database provides Weibull parameters for every location in Finland, thus allowing the simulation of scenarios with new non-measured locations.

Also, example cases for the applicability of the models were presented. These cases introduced different wind power generation structures. The results were assessed and compared against the aggregate wind power generation data from Finland. The example cases depicted possibilities these models provide for the power system operator for the planning of the power systems with high penetration of wind power.

References

[1] *Tuuliatlas, Finnish Wind Atlas Database*, Finnish Meteorological Institute. [Online]. Available: http://www.tuuliatlas.fi/en/index.html

[2] B. Klöckl, "Multivariate time series models applied to the assessment of energy storage in power systems," in *Proceedings 10th International Conference of Probabilistic Methods Applied to Power Systems*. IEEE, 2008, pp. 1–8.

[3] K. Xie, Y. Li, and W. Li, "Modelling wind speed dependence in system reliability assessment using copulas," *IET Renewable Power Generation*, vol. 6, no. 6, pp. 392–399, November 2012.

[4] K. Nystrom and J. Skoglund, "Univariate extreme value theory, garch and measures of risk," *Preprint, Swedbank*, September 2002, retrieved on June 2013. [Online]. Available: http://gloria-mundi.com/UploadFile/2010-2/knjsug.pdf

[5] R. Coić, J. Krstulović, and D. Jakus, "Simulation of aggregate wind farm short-term production variations," *Renewable Energy*, vol. 35, no. 11, pp. 2602–2609, May 2010.

[6] H. Louie, "Evaluation of bivariate archimedean and elliptical copulas to model wind power dependency structures," *Wind Energy*, November 2012.

[7] G. Papaefthymiou and D. Kurowicka, "Using copulas for modeling stochastic dependence in power system uncertainty analysis," *IEEE Transactions on Power Systems*, vol. 24, no. 1, pp. 40–49, February 2009.

[8] B. Stephen, S. J. Galloway, D. McMillian, D. C. Hill, and D. G. Infield, "A copula model of wind turbine performance," *Power Engineering Letter, IEEE Transactions on Power Systems*, vol. 26, no. 2, May 2011.

[9] X. He and P. Gong, "Measuring the coupled risks: A copula-based cvar model," *Journal of Computational and Applied Mathematics*, vol. 233, pp. 1066–1080, 2009.

[10] B. G. Brown, R. W. Katz, and A. H. Murphy, "Time series models to simulate and forecast wind speed and wind power," *Journal of Climate and Applied Meteorology*, vol. 23, May 1984.

[11] D. Villanueva, A. Feijóo, and J. L. Pazos, "Simulation of correlated wind speed data for economic dispatch evaluation," *IEEE Transactions on Sustainable Energy*, vol. 3, no. 1, January 2012.

[12] D. A. Bechrakis and P. D. Sparis, "Correlation of wind speed between neighboring measuring stations," *IEEE Transactions on Energy Conversion*, vol. 19, no. 2, June 2004.

[13] D. C. Hill, D. McMillan, K. R. W. Bell, and D. Infield, "Application of auto-regressive models to u.k. wind speed data for power system impact studies,"

IEEE Transactions on Sustainable Energy, vol. 3, no. 1, pp. 134–141, January 2012.

[14] B. Klöckl and G. Papaefthymiou, "Multivariate time series models for studies on stochastic generators in power systems," *Renewable Energy*, vol. 80, no. 3, pp. 265–276, March 2010.

[15] B. Deler and B. L. Nelson, "Modeling and generating multivariate time series with arbitrary marginals using a vector autoregressive technique," in *Proceedings of the 2001 Winter Simulation Conference*, 2001.

[16] G. Papaefthymiou and P. Pinson, "Modeling of spatial dependence in wind power forecast uncertainty," in *Proceedings 10th International Conference of Probabilistic Methods Applied to Power Systems*, Rincon, Puerto Rico, May 2008.

[17] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*, 2nd ed. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 13–31, 69–82, 387–398, 585–589.

[18] R. S. Pindyck and D. L. Rubinfeld, *Econometric Models and Economic Forecasts*, 4th ed. United States of America: The McGraw-Hill, 1998, pp. 122–128.

[19] "Wind power generation statistics of Finland," VTT, retrieved on Dec 2013. [Online]. Available: http://www.vtt.fi/proj/windenergystatistics/

[20] J. L. Myers and A. D. Well, *Research Design and Statistical Analysis*, 2nd ed. New Jersey, Mahwah: Lawrence Erlbaum, 2003, pp. 504–511.

[21] J. Rank, *Copulas: From theory to application in finance*. Haymarket House, London: Risk Books, a Division of Incisive Financial Publishing Ltd, 2007, pp. 3–34.

[22] C. Z. Mooney, *Monte Carlo Simulation*. Sage, 1997, pp. 1–12.

[23] L. E. Bantis, "Fit distributions to censored data," Matlab Central, September 2012, retrieved on June 2013. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/38226-fit-distributions-to-censored-data

[24] Z. Qin, W. Li, and X. Xiong, "Generation system reliability evaluation incorporating correlations of wind speeds with different distributions," *IEEE Transactions on Power Systems*, vol. 28, no. 1, Feb 2013.

[25] Z. Qin, W. Lin, and X. Xiong, "Estimating wind speed probability distribution using kernel density method," *Electric Power Systems Research*, vol. 81, no. 12, pp. 2139–2146, December 2011.

[26] A. Harvey, *The Econometric Analysis of Time Series*, 2nd ed. Cambridge, Massachusetts: The MIT Press, 1993, pp. 179–180.

[27] "Vestas V105-3.3 MW IEC IA wind turbine specifications," Vestas Wind Systems A/S, pp. 10–11, retrieved on Dec 2013. [Online]. Avail-

able: http://nozebra.ipapercms.dk/Vestas/Communication/Productbrochure/3MWbrochure/3MWProductBrochure/

A Bivariate Dependency Structures of the Presented Wind Speed Models

In this appendix, two bivariate dependency structures are presented as an additional information for Section 5.2.2 and Section 5.3.2.

Figure A1 presents the bivariate dependency structures specified by the transformed VAR model with time-dependent intercept term before the transformation from normal distribution to wind speeds. It can be observed, that the dependency structure resembles closely to the dependency structure of the standardized VAR model presented in Figure 8.

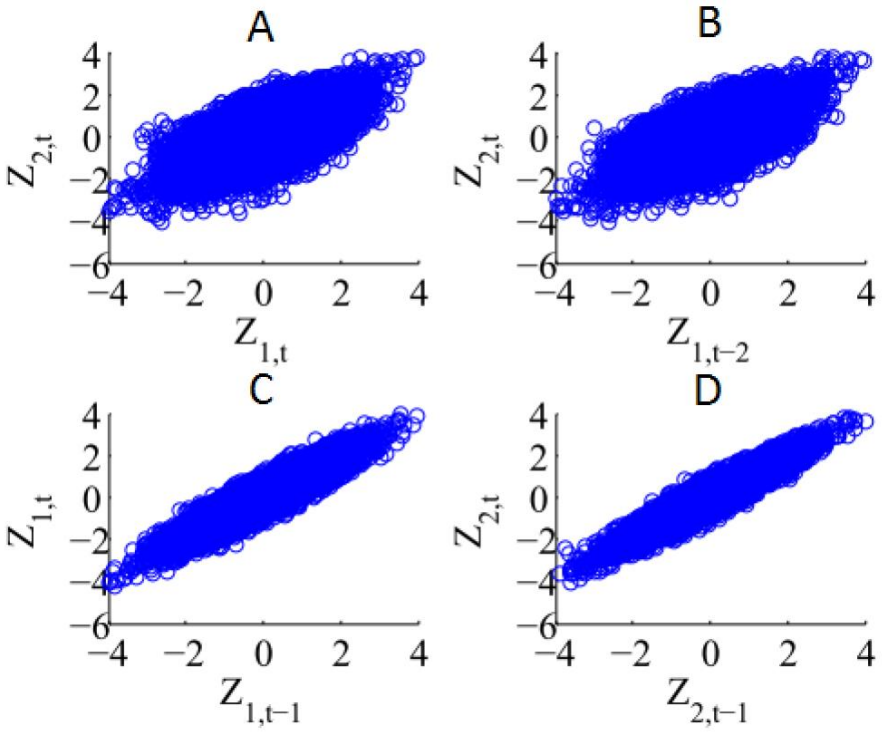

Figure A1: Four bivariate dependency structures specified by the transformed VAR model with time-dependent intercept term. In all four plots, both axes depict simulated normally distributed time series (in each plot subscript of $\boldsymbol{Z}$ specifies the location (1 or 2), and the moment (t, $t-1$ or $t-2$) for x- and y-axis) and the data is marked with blue circles. Plot A depicts the fully spatial case (dependency between two locations contemporaneously), plot B spatial and temporal (dependency between two locations with different moments t and $t-2$) and plots C and D both fully temporal cases (dependency between different moments t and $t-1$ in the same location. Simulations for figures were done with the high altitude data presented in Section 4.2.

Similarly, Figure A2 presents the bivariate dependency structures specified by the transformed ARC model with diurnal variations before the transformation from normal distribution to wind speeds. Respectively with the transformed VAR model,

it can be observed, that the dependency structure of the transformed ARC model also resembles closely to the dependency structure of the standardized VAR model presented in Figure 8.

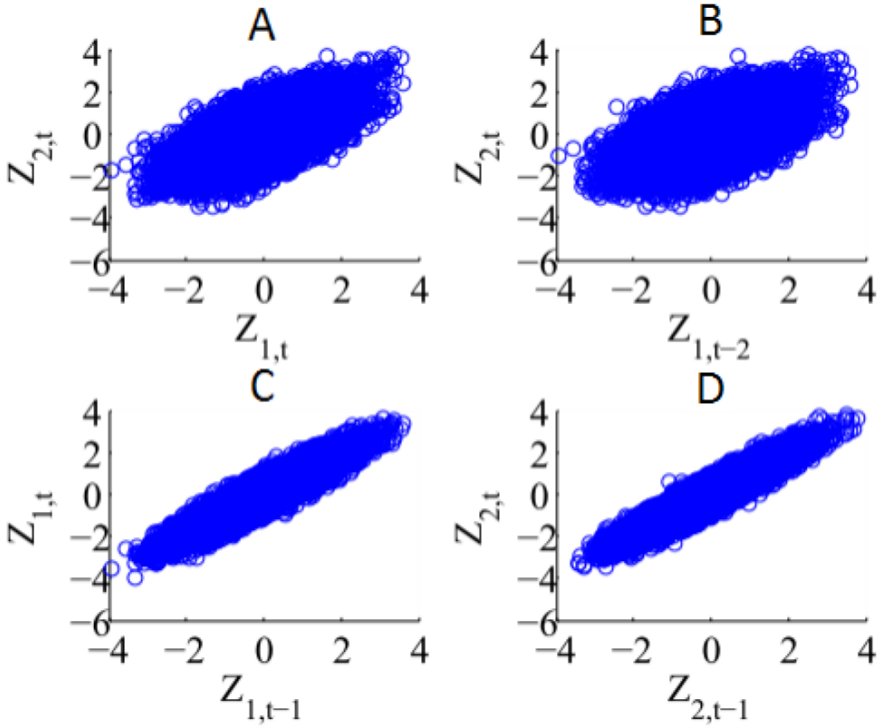

Figure A2: Four bivariate dependency structures specified by the transformed ARC model with diurnal variations. In all four plots, both axes depict simulated normally distributed time series (in each plot subscript of $\boldsymbol{Z}$ specifies the location (1 or 2), and the moment (t, $t-1$ or $t-2$) for x- and y-axis) and the data is marked with blue circles. Plot A depicts the fully spatial case (dependency between two locations contemporaneously), plot B spatial and temporal (dependency between two locations with different moments t and $t-2$) and plots C and D both fully temporal cases (dependency between different moments t and $t-1$ in the same location. Simulations for figures were done with the high altitude data presented in Section 4.2.